United States
Naval Postgraduate School

THESIS

A METHOD OF CLUSTER ANALYSIS

by

William Michael Cima

June 1970

United States
Naval Postgraduate School

THESIS

A METHOD OF CLUSTER ANALYSIS

by

William Michael Cima

June 1970

A Method of Cluster Analysis

by

William Michael Cima
Lieutenant (junior grade), United States Navy
B.S., United States Naval Academy, 1969

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 1970

ABSTRACT

A method of cluster analysis is presented in which points in n-dimensional space are analyzed through a subdivisive procedure. The points are orthogonally projected onto that line which maximizes their variance and the resulting point distribution is then analyzed with the use of a histogram. Wherever possible, divisions between conglomerates of points are made and each separate clump is subsequently analyzed. Ultimately adjacent groups are combined and analyzed through an analogous technique in an effort to re-unite any points which may have inadvertently deviated from the group with which they truly associate. The method is later refined to allow the detection of groups in several point dispersions which would have appeared as a single conglomeration under the original method. An example is given to illustrate the applicability of the procedure.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The author would like to express sincere thanks to Professor G. T. Howard for making available his time and talents which attributed much to the more technical aspects of this work. Special and greatly deserved thanks are extended to Professor Gary A. Tuck who did much to illuminate the subject matter. Through his authentic interest, patience and knowledge, he successfully led and directed the author toward the culmination of this work.

I. INTRODUCTION

A. EXPLANATION OF THE PROBLEM

It is frequently necessary to classify or distinguish between events or objects. The course of one's everyday life involves numerous automatic and casual classifications. For example, without an acute conscious awareness it is possible to discriminate between man and woman, bees and birds, or various colors. In effect, an entity is characterized through the outcome of the observations of several classification characteristics or variates. In cases where the classification is less obvious, each of these n classification variates may be treated as a dimension in n-dimensional space. The individual specimens on which the observations are being made may then be conceived of as points or n-dimensional vectors plotted on a multi-dimensional coordinate system whose axes are the scales of the specimen characteristics. Thus, every subject can be represented by one point in space.

Cluster analysis is a method of estimating the true groupings of these points. If one were observing characteristics of a sample from k different populations, p_1, p_2, $\ldots p_{i-1}$, p_i, p_{i+1}, $\ldots p_k$, he would then want to divide n-dimensional space into k mutually exclusive and collectively exhaustive regions, r_1, $r_2, \ldots . r_{i-1}$, r_i, r_{i+1}, $\ldots$, r_k, with the rule or procedure of assigning an individual to p_i if he belongs to r_i. In other words, cluster analysis

circumscribes the k areas of high point density in hyperspace with multi-dimensional boundaries which define divisions between the k populations.

A procedure which could accomplish this has wide applicability. For example, the dispersion of human chromosomes as they appear on a photograph of mitosis involves a two-dimensional array of points; the dispersion of stars involves a three-dimensional array: bacterial classification could involve an n-dimensional array as would any taxonomical problem in which n characters of m species are treated as variates and represented by m points in n-dimensional space. Information on the true grouping of the species could be obtained from the point grouping as seen through cluster analysis. Thus, most investigations into the classifications of objects or dispersion of things require the elucidation of spatial groupings.

B. GENERAL

There are two distinct methods of attacking the cluster analysis problem. The first of these is the agglomerative method which starts with a single point as the "group" and adds to it others which satisfy certain criteria. A group is formed when no new points can be added to the existing conglomeration; another starting point is then chosen and the process iterates to determine the second group. This technique is continued until all points are allocated to one of the k groups. Thus, groups are built up from the individual points and in this manner n-dimensional space is

divided into k mutually exclusive and collectively exhaustive regions. The other method of approach employs subdivisive techniques which arrive at an ultimate partition of n-dimensional space through an initial partition, or possibly several initial partitions, of n-dimensional space and subsequent divisions of each of those partial spaces.

C. HISTORY

The area of cluster analysis has received ever increasing attention over the last four decades by those interested in classification methods. Four of the more renowned procedures are briefly reviewed; if it is desired to learn more concerning any of these methods the list of references will be helpful in directing the reader to sources of greater detail. In 1933, Hotelling devised a principle component analysis method of grouping points which involved the successive elimination of dimensions. Points are projected orthogonally from a multi-dimensional space into a space of fewer dimensions which retains "maximum information." Said differently, those characteristics having the least amount of variability among their observations are eliminated and the space reduced. This process is continued until the points are finally in an observable space.

The weighted mean pair method developed by Sokal and Michener (1958) and altered by Rogers (1959) has been applied to entomological problems. The procedure, again updated in 1963 by Sokal and Sneath, involves operations performed with an initially constructed m by m symmetric

matrix, where m is the number of points. The resemblance between two specimens is the proportion of the number of characteristics measured which they have in common. Thus, the larger this similarity ratio, the more alike are the specimens. Then a distance, d_{ij}, between any two points, c_i and c_j, $i, j = i, 2,...m$, is defined as $d_{ij} = -\log s_{ij}$, where s_{ij} is the similarity ratio between points c_i and c_j. It is interesting to notice that as s_{ij} approaches 1, d_{ij} approaches 0. Thus, the symmetric similarity matrix (d_{ij}) is in effect a mileage chart in semi-metric space, showing distance between any two points c_i and c_j. Then an overall number, H_i, is computed for each point c_i, where $H_i = \sum_j d_{ij}$, $j = 1, 2,...m$, and $i = j$. The point possessing the smallest H_i value is designated as the prime node and becomes the first member of the group. Without loss of generality it can be labeled c_1. The point closest to c_1, c_i, is then added to the group if its similarity coefficient satisfies certain specified criteria. With this, the first and i^{th} rows and columns of the matrix are deleted, one new row and column representing the combination of the similarity coefficients of c_1 and c_i added, and the process repeated for the $n-1$ order matrix. This method continues until no new points have similarity coefficients which satisfy the group entry criterions. Then a smallest H_i value is computed for those points which were not admitted to the group and the process iterates to form multi-groupings until all points are allocated to a group.

In 1959, Williams and Lambert concerned themselves with the case of an $n \times m$ data matrix X, where n is the number of points and m is the number of variates, consisting entirely of presence or absence data (1 or 0 respectively). Originally they applied their method to taxonomic problems in ecology, where the variates were the different plant characteristics present or absent in m individuals. The m points are divided into two subsets on the basis of the variate k which "best" separates them (in a well-defined sense): One set being those that contain k, the other being those that do not. If it is feasible to divide the groups with respect to any of the remaining variates, the appropriate dimension is chosen and the process continued.

Edwards and Cavalli-Sforza (1965) developed an accurate but highly tedious subdivisive procedure. Noting that the best division between clusters would be that which resulted in the two clusters being as dense as possible, they proceed to separate the points into two groups through every possible division of points. They then choose that division which maximizes the between group variance and minimizes the within group variance. The procedure is repeated for each group with a weighting factor, the associated between clusters sum of squares, to describe the importance of each division. Note that the drawback to this method lies in the computational labor of examining all splits. The authors admit that $(n-1)^2\ 2^{n-1}$ seconds are required on a computer

with a five microsecond access time; that is, twenty-one points require 100 hours and forty-one points require 54,000 years.

Presented in the following pages is a method of cluster analysis. An example is given to show how the procedure is of assistance in analyzing problems of a taxonomical nature.

II. PROCEDURE

A. BACKGROUND

The procedure presented in this paper employs subdivisive techniques. The method involves the orthogonal projection of points in n-dimensional space onto a one-dimensional line and an analysis of the resulting point distribution. In order to facilitate the group detection process, it seems desirable to determine that line which provides maximum separation between the group projections. The difficulty in doing this emanates from not knowing which points associate to form groups. A second impediment is that the meaning of maximum separation between groups is not clear. The concept of maximizing separation between groups is subject to several different interpretations, two of which follow: the line providing maximum separations between groups is obtained when (1) the sum of the distance between the closest points of all adjacent groups is a maximum or (2) any two adjacent groups have a maximum separation, that is, the separation between the projections of two adjacent groups is greater than the separation between the projections of any other adjacent groups on any line. Figure 1 may be helpful in clarifying this distinction:

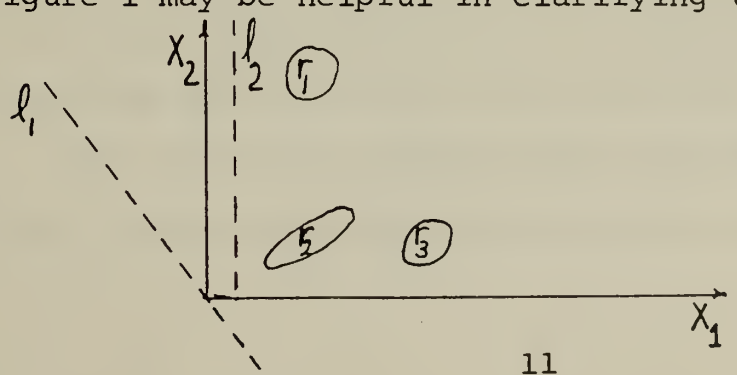


Figure 1

The line on which the sum of the distances between the closest points of all adjacent group projections is a maximum is ℓ_1. That is, the sum of the distance between groups one and two plus that between groups two and three is greatest when projected on line ℓ_1. It thus corresponds to the first interpretation. The maximum separation between the projections of two groups, r_1 and r_2 is provided by ℓ_2. There is no line on which the projections of any two adjacent groups are separated by a distance greater than that between groups r_1 and r_2 on line ℓ_2. Thus, ℓ_2 provides maximum separation with respect to the second interpretation.

Inherent to subdivisive methods is the possibility of erring by assigning a point to a group with which it does not conglomerate. It is, therefore, desirable to make divisions between groups in a manner which minimizes the probability of introducing such error. Thus, in every case it is advantageous to work with lines analogous to ℓ_2 because it provides a maximum amount of separation between two high density point areas which are to be divided into two groups. In this manner, it reduces the probability of erring by assigning points to the wrong group. However, certain point dispersions which result in well-defined groups enhance the utility of lines analogous to ℓ_1, which has the advantage of dictating more divisions per iteration.

One line which maximizes the separation between groups under the first interpretation is that line which maximizes

the between group variance and minimizes the within group variance. However, inadequate knowledge concerning those vectors which conglomerate or which points associate to form groups makes this method non-functional computationally; in this form it involves the analysis of all possible divisions of points into groups and the subsequent projection of each combination onto the most suitable line. The computational labor of such an analysis rivals that of the method of Edwards and Cavalli-Sforza mentioned earlier.

B. THE APPROPRIATE LINE

The line which is used is not as discriminating as that which maximizes the between group variance and minimizes the within group variance, but is computationally more accessible. The basic analytical tool is chosen to be that one-dimensional ray on which the variance of the orthogonally projected points is a maximum. Several point distributions can be formed for which that line which maximizes the variance of the projected points does not coincide with or even lie near that line which maximizes the separation between groups. However, it does appear to be a reasonable line in that in most cases it provides a point distribution which is very auspicious to group detection in later analysis. The problems arising when this line is not favorable are reckoned with later. The problem of finding a line which maximizes the variance of the orthogonally projected points can be formulated as a non-linear programming problem. The objective function is to maximize the variance

of the projected points on a line given by the direction of u, a unit vector.

The problem is:

$$\underset{u}{\text{Max}}\ S^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$$

Subject to: $u^t u = 1$,

where:

S^2 = sample variance

m = number of points

$$x_i = \sum_{j=1}^{n} c_{ij}\, u_j$$

where:

n = the number of dimensions or variables

$u^t = (u_1, u_2, \ldots, u_n)$ = unit vector

$c_{ij} = (c_i, c_{i2}, \ldots, c_{in})$ = coordinates of the i^{th} point in n dimensions

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij}\, u_j).$$

The problem now becomes:

$$\text{Max}\ S^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} c_{ij}\, u_j \right) - \left(\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\, u_j \right) \right]^2$$

subject to $u^t u = 1$.

The Lagrangian is then:

$$F(u_1, \ldots, u_n, \lambda) = S^2 - \lambda(u^t u - 1)$$

but $u^t u = \sum_{j=1}^{n} u_j^2$,

which implies

$$F(u_1, \ldots, u_n, \lambda) = s^2 - \lambda\left(\sum_{j=1}^{n} u_j^2 - 1\right),$$

and substituting in the expression for s^2 gives

$$F(u_1, \ldots, u_n, \lambda) = \frac{1}{m-1} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} c_{ij} u_j \right) - \left(\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} u_j \right) \right]^2 - \lambda \left(\sum_{j=1}^{n} u_j^2 - 1 \right) .$$

Setting the partial derivative of the Lagrangian with respect to each variable equal to zero gives necessary conditions from which the optimal vector, u, can be found. For the k^{th} component of u:

$$0 = \frac{\partial F}{\partial u_k} = \frac{2}{m-1} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} c_{ij} u_j \right) - \left(\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} u_j \right) \right] \cdot$$

$$\left[c_{ik} - \frac{1}{m} \sum_{i=1}^{m} c_{ik} \right] - 2 \lambda u_k .$$

Letting $c_{ik} - \frac{1}{m} \sum_{i=1}^{m} c_{ik} = B_{ik}$ and rewriting gives:

$$0 = \frac{2}{m-1} \sum_{i-1}^{m} \left[\left(\sum_{j-1}^{n} c_{ij}\, u_j \right) \left(B_{ik} \right) - \left(\frac{1}{m} \sum_{i-1}^{m} \sum_{j-1}^{n} c_{ij}\, u_j \right) \left(B_{ik} \right) \right] - 2\, \lambda u_k .$$

Distributing the first summation over the terms:

$$0 = \frac{2}{m-1} \sum_{i-1}^{m} \left[\sum_{j-1}^{n} c_{ij}\, u_j \right] B_{ik} - \frac{2}{m(m-1)} \sum_{i-1}^{m} \left[\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\, u_j \right] B_{ik} - 2\, \lambda u_k .$$

Rearranging within each term:

$$0 = \sum_{i=1}^{m} \frac{2\, B_{ik}}{m-1} \sum_{j=1}^{n} c_{ij}\, u_j - \left[\sum_{i=1}^{m} \frac{2\, B_{ik}}{m(m-1)} \right] \left[\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\, u_j \right] - 2\, \lambda u_k .$$

The second term is identically zero since

$$\sum_{i=1}^{m} B_{ik} = \sum_{i=1}^{m} \left(c_{ik} - 1/m \sum_{i=1}^{m} c_{ik} \right) = 0 ,$$

which shows that

$$\sum_{i=1}^{m} \frac{2\,B_{ik}}{m(m-1)} \left[m\bar{x} \right] = 0 .$$

Therefore,

$$\frac{\partial F}{\partial u_k} = 0 = \sum_{i=1}^{m} \frac{2\,B_{ik}}{m-1} \sum_{j=1}^{n} c_{ij}\, u_j - 2\,\lambda\, u_k .$$

Setting $\frac{2\,B_{ik}}{m-1} = D_{ik}$ yields

(1) $$0 = \sum_{i=1}^{m} D_{ik} \sum c_{ij}\, u_j - 2\,\lambda\, u_k \text{ , } h = 1, \ldots n, \text{ as}$$

a necessary condition. Differentiating the Lagrangian with respect to the Lagrange multiplier, λ , and setting the derivative equal to zero yields the constraint

(2) $$u_1^{\,2} + \ldots + u_n^{\,2} = 1 .$$

Thus, u is found through the solution of a system of n+1 equations in n+1 unknowns, $u_1, u_2, \ldots, u_n$. There are n equations of the form (1) and there is one equation of form (2). It is important to notice that setting the first partial derivatives of the Lagrangian equal to zero is a necessary but not a sufficient condition for a maximum. In fact, on rare occasions when the rank of the Jacobian of the constraint equations is less than m, the number of

constraints, these conditions are not even necessary. That is, for a set of constraints possessing a singular Jacobian, there may exist solutions which maximize or minimize the objective functions which will not be found using the first partial derivative of the Lagrangian technique. This will be ignored here, however. Since there is no guarantee that a solution to the above system of equations maximizes the objective function, it is imperative that all solutions be obtained. The unit vector which maximizes the variance of the projected points will be one of those solutions and can be identified as the maximum through its substitution into the objective function.

C. UNDERLYING THEORY

Before delving further into the details of the procedure, one assumption can be made from which two important observations follow; it is hypothesized that the n-dimensional hypervolume spanned by the points of any group is nearly a hyperellipsoid whose lengths along the axes are real numbers greater than or equal to zero. This assumption seems consistent with expectations since a hyperellipsoid is the shape of the hypervolume that the points will most likely span. Furthermore, it imposes virtually no constraint since it permits the spatial dispersion of points of any group to generate a multi-dimensional hypervolume of many forms, from a hyperball (solid hypersphere) to a straight line segment.

The assumption that the hypervolume spanned by the points is a hyperellipsoid results in the first observation --that a curve-fitted histogram plot of the distribution of points of any group projected on a line has a bell-like structure similar to that of the normal density; furthermore, the effect is especially inherent to groups of high point density. This phenomenon can be reasoned as follows: ℓ_1, will represent the line onto which the points are to be orthogonally projected and the two-dimensional case will be considered. A sufficient number of equidistant parallel lines, w_1, w_2,..., w_j, can be constructed which completely bound and divide the entire group. The value of j would depend upon the distance between lines. In Figure 2 below, it can be observed that the expected number of points of the

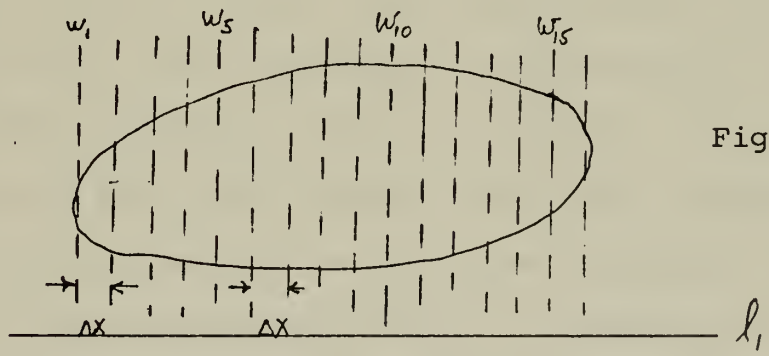


Figure 2

group contained between w_8 and w_9 is greater than that contained between w_3 and w_4, which in turn is greater than that between w_1 and w_2. Hence, the curve-fitted histogram can be expected to exhibit a bell-like structure. Figure 3 indicates that the orientation of the group with respect to the line is free to vary without altering this result. Now generalizing this two-dimensional analysis for the n-dimensional case, ℓ_1, can again represent

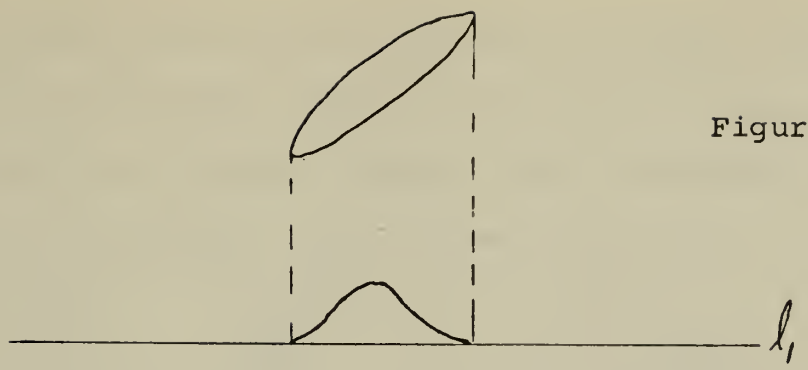


Figure 3

the line onto which the points are to be projected. Then, as before, it is possible to construct a sufficient number of equidistant parallel (n-1)-dimensional hyperplanes, $w_1, w_2, \ldots, w_j$, to completely bound and divide the hyper-ellipsoid. Then, intuitively, the expected number of points contained between two adjacent hyperplanes passing through a "thick" portion of the ellipsoid (near its geometric center) is greater than that contained between two hyperplanes in a "thinner" portion. Hence, the curve-fitted histogram would possess a bell-like structure, an effect which is more salient for increasing density of points. An interesting extension of this observation is that the normal-like shape becomes more pronounced as the hyper-ellipsoid tends to a hyperball and less pronounced as it elongates; in fact, the bell-like structure is not discernible for a hyperellipsoid in the form of a straight line. The concept of ascertaining information concerning the group shape through the analysis of point distributions on the line will later be shown to be of value.

The second observation follows from the first: a histogram of the projections of two or more groups superimposed on the line will vary among bell-like, skewed bell, flattened

bell, or any combination of bell-like forms. This observation is intuitively appealing and a few illustrations will show that it is indeed logical. The histogram can be expected to contain local maxima and minima.

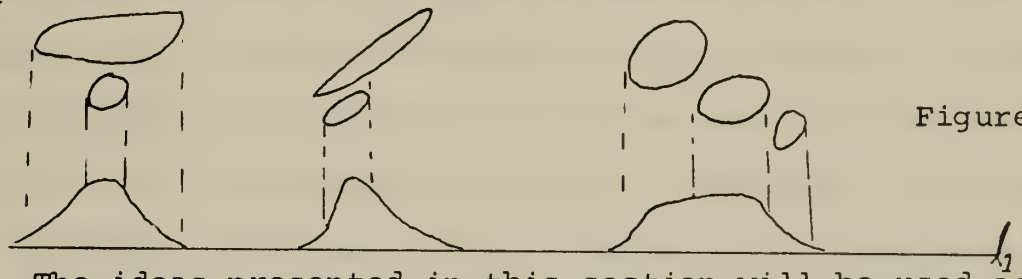


Figure 4

The ideas presented in this section will be used as keystones in those portions of the procedure concerned with analysis of histograms.

D. THE ANALYSIS OF PROJECTED POINTS

As stated earlier, the procedure being developed involves the projection of points on a line and the subsequent examination of the distribution of projected points. The dispersion of points on the line is analyzed through the use of a histogram. Divisions are made between maxima and each segment again analyzed through the use of a new line. Ultimately, adjacent groups are combined and analyzed to correct the possible error that too many divisions were made.

Assuming that the line which maximizes the variance of the points has been found using the non-linear programming method shown previously, the next step is to form a histogram. Fundamental to its construction is a decision concerning the interval length between the equidistant parallel (n-1)-dimensional hyperplanes. It has been determined that, in most cases, dividing the line length spanned

by the two end points into m/2 equal intervals, where m is the number of points under consideration, appears to be a reasonable starting interval length. Then if this does not suffice, the interval length yielding more well-defined maxima can be determined through experimentation if desired.

Assuming the efforts in histogram construction are successful, its analysis can now be undertaken. If the plot is unimodal this phase of the procedure is terminated and it can be concluded, for the moment, that all those points comprise one group. As a matter of definition, a plot is considered to be unimodal if there do not exist two separate peaks with a point between them less than eight-tenths (.8) the height of the lesser peak, having zero slope and positive second derivative. Eight-tenths is chosen through the following reasoning: it is advantageous to divide the space into a maximum number of regions to gain information concerning the spatial grouping of points. As the number of groups analyzed simultaneously decreases, the accuracy and informational yield of the analysis increases. The fact that groups will later be combined alleviates any fear of error emanating from splitting conglomerates of points which are actually members of the same fraternity. Thus, it is beneficial to the discrimination process to impose and exercise a rather ruthless division rule, that is, one which dictates divisions readily. On the other hand, an expected result of the non-uniformity of point density is that the histogram of a group contains slight dips and rises

superimposed on an overall bell-like form. Therefore, the ruthlessness in division must be reduced to account for a jagged histogram form. Thus, the cut-off is arbitrarily set to be a dip of twenty percent (20%) the height of the lesser peak. The following illustrations may be helpful in clarifying this concept:

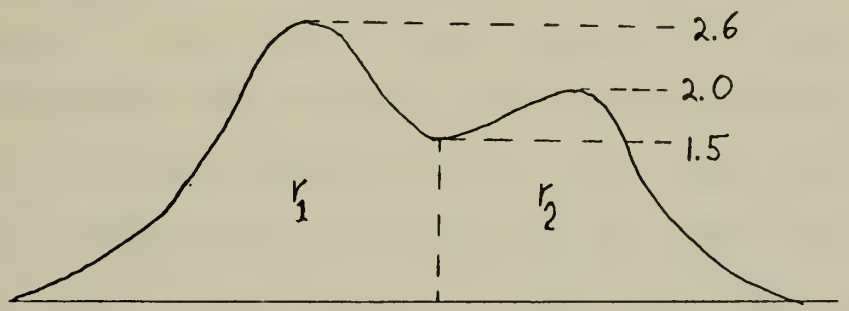


Figure 5

In the above situation, the value of the zero slope point between the two local maxima is sufficiently low to allow a partition of the space into two regions. The second curve permits a three-way partition of n-dimensional space.

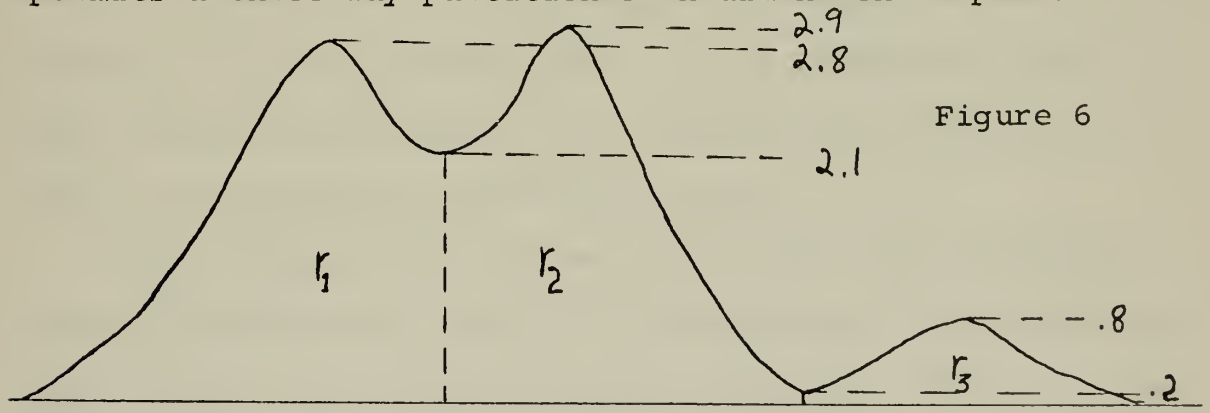


Figure 6

Only one division can be made in the final diagram.

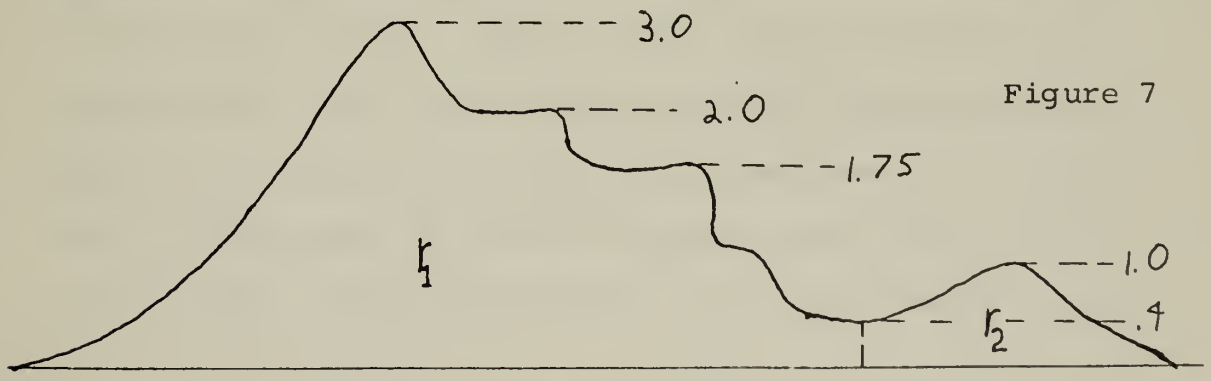


Figure 7

After determining the number of maxima or groups the histrogram portrays, the points between any two adjacent peaks must be separated between them. The decision rule for this is to make a division between the maxima at the point c_i where the curve-fitted histogram takes on zero slope and has a positive second derivative as indicated in Figures 5 through 7. Then all of those points to the left of c_i are allocated to the maximum on the left and those to the right of c_i to the maximum on the right. Note that this allocation is made with knowledge that for those cases in which the value of the histogram at the zero slope, positive second derivative point deviates from zero, i.e., for overlapping projections of groups on the line, the method will cause some points to be allocated to a group with which they do not cluster. This inaccuracy is accepted for the present because of the possibility that in the subsequent analysis, those points allocated to the inappropriate group will be split off and later correctly combined.

Then, assuming all of the points have been allocated among several maxima, that is, n-dimensional space has been partitioned into several mutually exclusive and collectively exhaustive regions, the process is continued for each of the regions separately. That is, the entire procedure for finding the best line and analyzing the distribution of points is iterated for each of the regions individually. This is continued for each subregion until none of the regions will subdivide further, or in other words, until

there does not exist a histogram of any of the regions which contains more than one maximum. A summary of this divisive procedure for a typical distribution of points could be represented schematically as follows:

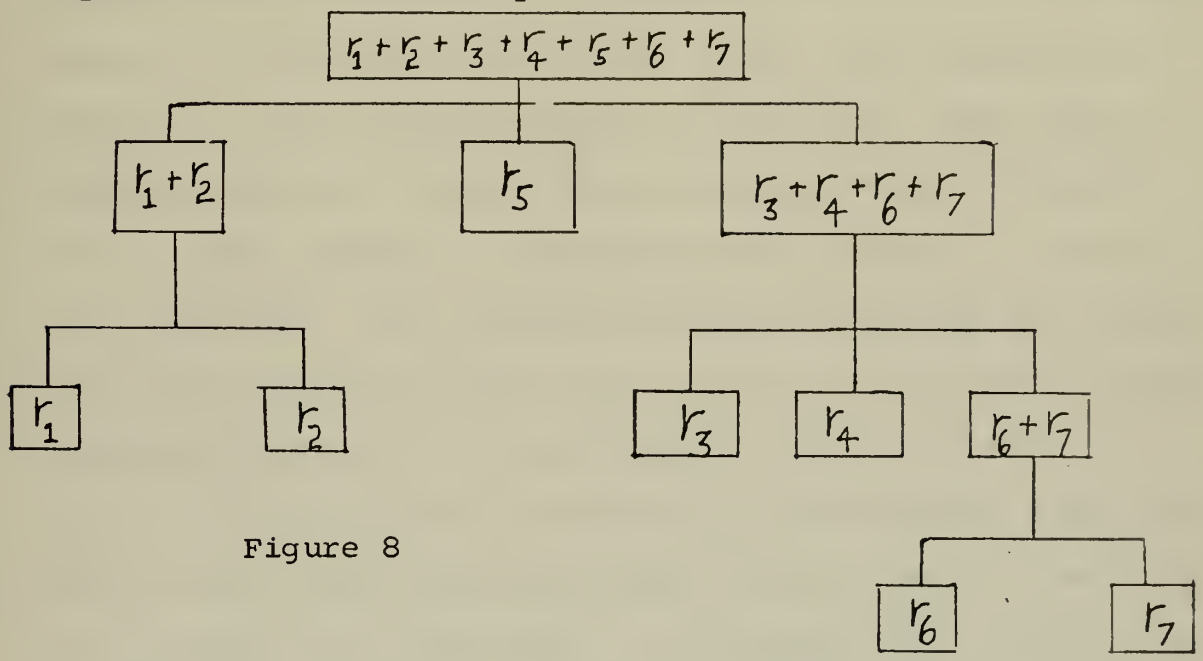


Figure 8

In Figure 8 the procedure has divided n-dimensional space into seven regions. The original maximum variance line contained three distinct maxima and n-dimensional space was partitioned into three regions, $r_1 + r_2$, r_5, and $r_3 + r_4 + r_6 + r_7$. The maximum variance line for $r_1 + r_2$ and $r_6 + r_7$ showed two maxima and that for $r_3 + r_4 + r_6 + r_7$ contained three. Regions r_1, r_2, r_3, r_4, r_5, r_6, r_7 were unable to further divide because each had unimodal histograms of their individual dispersion of points when projected on the maximum variance line. The result is that n-dimensional space has been partitioned into seven mutually exclusive and collectively exhaustive regions.

E. CORRECTING THE ERROR

As noted previously, in the process of separating points between peaks based solely on the zero slope, positive second derivative criteria, points may have been split from group r_i with which they truly cluster and included in group r_j. If this phenomenon has occurred, there then exist two possible routes these points may have followed: (1) in later analysis of histograms for group r_j or any of its subgroups, the foreign points were observed to cluster and form a separate peak; they were then split away from the remaining points of r_j and formed a separate group or (2) due to insufficient clustering, no new maxima containing these points were formed and they were not split from group r_j or any of its subgroups. A procedure for resolving the problem generated by the second alternative has not been developed and must be absorbed as risk. However, the probability of taking the first route is much less than that of taking the second, a result of the aggressive dividing rule.

The implication of the first alternative is that n-dimensional space has been partitioned into too many regions. By combining some of the regions this error can be rectified; that is, adjacent groups must be combined to reduce the number of regions. For definitional purposes those groups adjacent to group r_i will be (1) that group whose centroid is the least number of group r_i's standard deviations from group r_i's centroid and (2) that group

containing the point which is closest to any point of group r_i's, or in other words, that group whose hypersurface is closest to group r_i's hypersurface. The standard deviation in n-dimensions is calculated solely on the basis of distance from the mean and is thus not a directional quantity. The sample variance is

$$s^2 = \frac{1}{m-1} \sum_{i=1}^{m} (c_i - \bar{c})^2$$

and the standard deviation is $\sqrt{s^2}$,

where:

m = the number of points

c_i = the coordinates of the ith point = $(c_{i1}, c_{i2}, \ldots, c_{in})$

n = the number of dimensions

$\bar{c}$ = the coordinates of the mean point = $(\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_n)$

$$= \frac{1}{m} \left(\sum_{i=1}^{m} c_{i1}, \sum_{i=1}^{m} c_{i2}, \ldots, \sum_{i=1}^{m} c_{in} \right)$$

Thus the distance of any point, c_i, from the mean is

$$d_i = \left[\left(c_{i1} - \frac{1}{m} \sum_{i=1}^{m} c_{i1} \right)^2 + \left(c_{i2} - \frac{1}{m} \sum_{j=1}^{m} c_{i2} \right)^2 + \ldots + \left(c_{in} - \frac{1}{m} \sum_{i=1}^{m} c_{in} \right)^2 \right]^{\frac{1}{2}}$$

Both definitions of adjacent groups are applied in Figure 9.

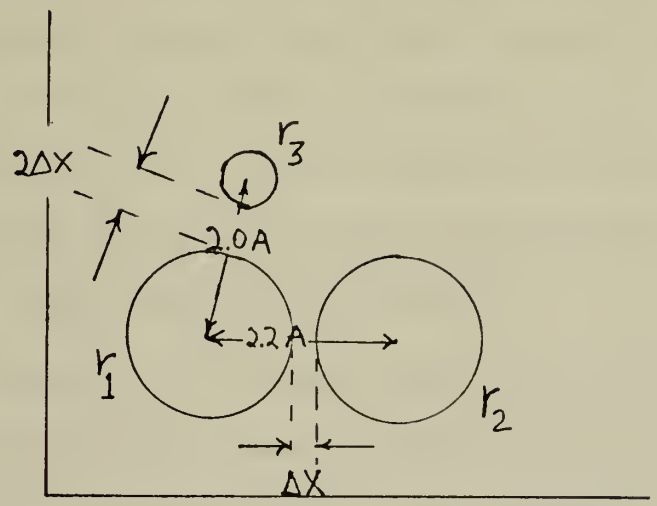


Figure 9

In Figure 9 group r_3's centroid is only two of group r_1's standard deviations from group r_1's centroid and is thus adjacent. Group r_2 is also adjacent to r_1 since its surface is closest to r_1's.

Thus, for group r_i with each of its adjacent groups individually in n-dimensional space, the optimal line is determined and the histogram analyzed. The rule for making divisions becomes much more stringent than in the earlier phases of analysis in that partitions of space are not made as readily. The combination of points is separated only if there exists a point between two maxima such that the slope of the curve-fitted histogram is zero, and the value of the curve also equals zero. The reason for this change in policy is that to err by making too few divisions, which appears to be unlikely, is seemingly more acceptable than to err by making too many divisions. This phase is continued until none of the groups will combine with any of its

neighboring groups to form one clump. A typical combination process of the earlier illustrated partitioning of n-dimensional space into seven groups is represented by Figure 10. Initially there are seven groups, r_1, r_2, ..., r_7. The symbols in each box represent the group and those immediately below the box represent the adjacent groups of the group designated in the box.

First r_1 is taken with each of its adjacent groups and it is found that it combines with r_2 but not r_7. In the second step it is determined that $r_1 + r_2$ will not combine with its adjacent group r_4. When r_3 is examined with each of its adjacent groups it is found that it does not unite with r_5 but will with $r_1 + r_2$. In the third step it can be seen that group $r_1 + r_2 + r_3$ will not combine with its adjacent group, r_5. Finally it is found that r_5 combines with r_6. In the last step, r_5 and r_6 are united with r_7, resulting in three final groups, $r_1 + r_2 + r_3$, r_4 and $r_5 + r_6 + r_7$.

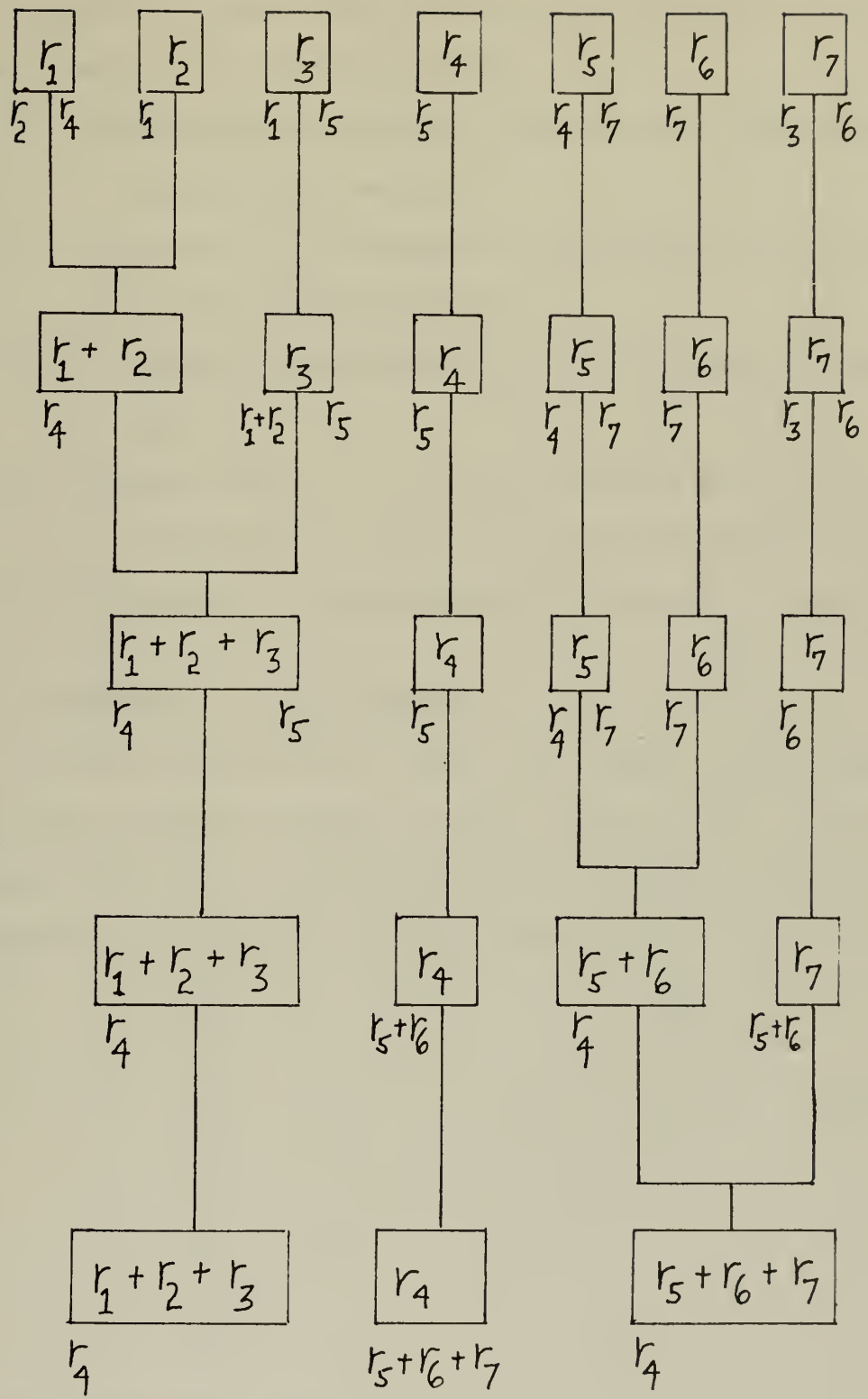


Figure 10

Thus far cluster analysis under this method may be summarized in the following steps.

(1) Determine the optimal line using the non-linear programming method.

(2) Construct a histogram representing the distribution of the projected points.

(3) Analyze the histogram and form a group for each peak.

(4) Repeat Steps 1, 2, and 3 for each mode and its associated group until none will subdivide.

(5) Analyze the combinations of adjacent groups.

F. REFINEMENT OF THE PROCEDURE

As mentioned earlier, there are cases in which the line providing maximum variance of the points is not favorable to group detection. One such case involves two side-by-side elongated ellipsoids. Consider Figure 11.

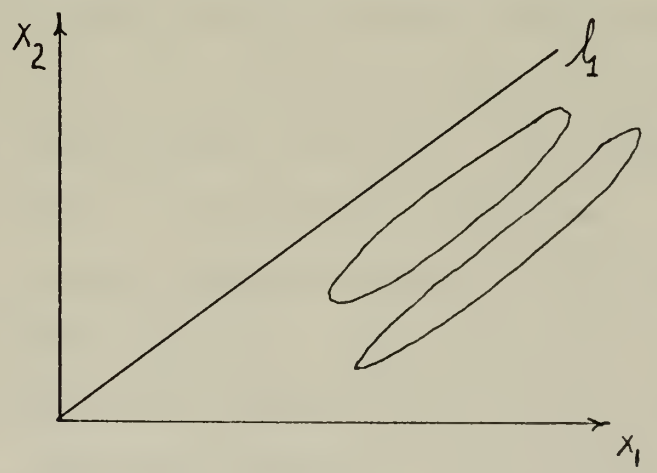


Figure 11

In this instance, the points would be projected on line l_1 and only one group would be detected. Similarly, in n-dimensional space the same problem would be posed, that is,

the maximum variance line would make it possible to detect only one group. It is, therefore, necessary to amend the procedure in a manner which can resolve the elongated ellipsoid case and many of its variations. In most cases it can be seen that there does exist a line onto which the points could be projected which would result in a histogram containing two maxima and hence, the detection of two groups. Figure 12 shows that this is indeed the situation:

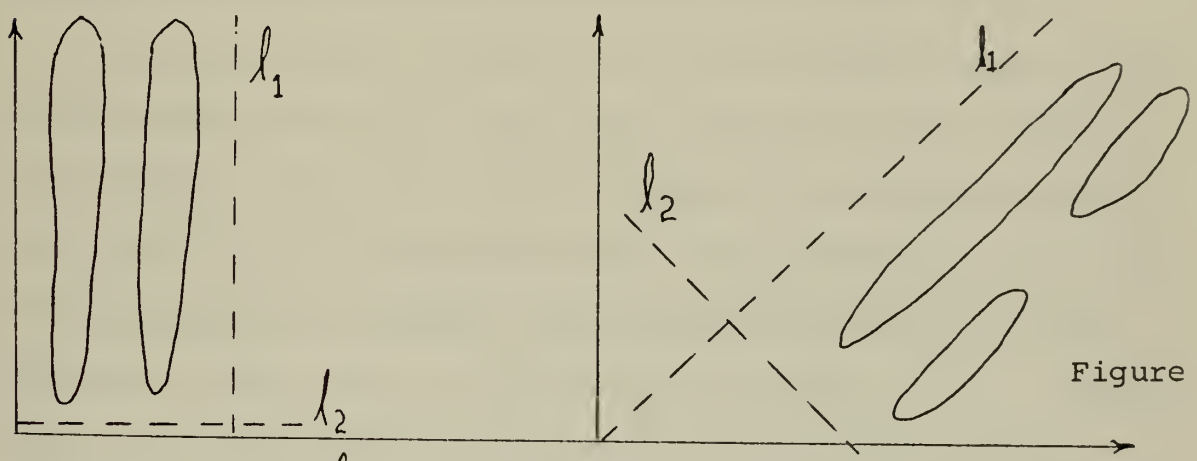


Figure 12

In each diagram, l_1 represents the maximum variance line and l_2 a line which facilitates the detection of two groups. It is interesting to notice that in each case l_2 is perpendicular to l_1. For the spatial dispersions of points under consideration, that is, for elongated ellipsoids, this result is intuitively appealing.

However, as more characteristics are observed and the space is extended beyond two dimensions, the number of family of lines directionally perpendicular to the maximum variance line, l_1, increases from one in two space to infinity when the dimension of the space is three or more. In the two dimensional case it can be reasoned that there is only one line family of parallel lines perpendicular to line l_1.

For the purpose of analyzing point distributions and to be consistent with the non-linear programming solution for the unit vector in the direction of ℓ_1, it is reasonable to specify that all lines onto which points are to be projected must pass through the origin. In essence, it is the slope of the line, not its axes intercepts, which is of importance. Similarly, in n-dimensional space the direction or spatial orientation of the line is the only concern.

In the following argument it can be assumed without loss of generality that all lines under discussion pass through the origin. Also ℓ_1 will designate the maximum variance line and ℓ_2 will designate the line perpendicular to ℓ_1, which successfully handles the elongated ellipsoid. The problem is thus one of ascertaining the line, ℓ_2, orthogonal to the maximum variance line which could better discriminate between areas of high point density. It is, therefore, desirable to find a line on which a histogram of the distribution of the projected points will be unimodal if only one group exists and contain two local maxima if two exist. This line, ℓ_2, is that line which is perpendicular to ℓ_1 and on which the variance of the projected points is a relative maximum. That is, considering Figure 12, ℓ_2 provides a maximum variance between points relative to all other lines perpendicular to the absolute maximum variance line ℓ_1, of which there are none. In the three-dimensional analog it can again be reasoned that the line which will facilitate the detection of two groups, if two

exist, is analogous to ℓ_2. It is again that line which is perpendicular to the absolute maximum variance line, ℓ_1, and is itself a relative maximum variance line, relative to all those lines perpendicular to ℓ_1. As the dimension of the space is increased from three to n-dimensions, the line ℓ_2 which is of interest is expected to possess the same characteristics.

The problem of finding such a line can be formulated as a non-linear programming problem of much the same structure as the original method. The only mutation is the addition of a constraint which specifies that the unit vector in the direction of ℓ_2 must be perpendicular to that in the direction of ℓ_1. Hence, the dot product of the two unit vectors is zero.

The non-linear programming problem is thus:

$$\text{Max } S^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$$

subject to

$$w^t w = 1$$

$$w^t u = 0$$

where: w is a unit vector in the direction of ℓ_2

u is a unit vector in the direction of ℓ_1

(note that all of the coefficients of u are known at this point) and the remaining symbols are unchanged in interpretation from the original programming problem.

The problem now becomes

$$\text{Max } S^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n} c_{ij} w_j \right) - \left(\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij} w_j) \right) \right]^2$$

subject to

$$w^t w = 1$$

$$w^t u = 0 .$$

The Lagrangian is thus

$$F(w_1, w_2, \ldots, w_n, \lambda_1, \lambda_2) =$$

$$s^2 - \lambda_1 (w^t w - 1) - \lambda_2 (w^t u) =$$

$$\frac{1}{m-1} \sum_{i-1}^{m} \left[\left(\sum_{j=1}^{n} c_{ij} w_j \right) - \left(\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} w_j \right) \right]^2$$

$$- \lambda_1 (w^t w - 1) - \lambda_2 (w^t u) .$$

As before the partial of the Lagrangian with respect to each variable can be set equal to zero to find the optional vector w.

Taking the partial with respect to the k^{th} component of w yields

$$0 = \frac{\partial F}{\partial w_k} = \frac{\partial}{\partial w_k}\left[s^2 - \lambda_1 (w^t w - 1) \right] - \frac{\partial}{\partial w_k} (\lambda_2 w^t u).$$

The first term is exactly the same as in the original problem with u replaced by w and thus the equation reduces to

$$0 = \sum_{i=1}^{m} D_{ik} \sum_{j=1}^{n} c_{ij} w_j - 2\lambda_1 w_k - \frac{\partial}{\partial w_k} (\lambda_2 w^t u)$$

where $D_{ik} = \frac{2B_{ik}}{m-1}$ and $B_{ik} = c_{ik} - \frac{1}{m} \sum_{i=1}^{m} c_{ik}$.

Differentiating the last term with respect to w_k yields only one term, that component of u which is multiplied by w_k.

The derivative of F with respect w_k reduces to

$$(3)\quad 0 = \sum_{i=1}^{m} D_{ik} \sum_{j=1}^{n} c_{ij} w_j - \lambda_1 w_k - \lambda_2 u_k, \quad k = 1, \ldots, n.$$

There will be n equations of this form, where n is the number of dimensions. Differentiating with respect to the Lagrange multipliers and setting the derivative equal to zero yields the constraints:

$$(4)\qquad w^t w = 1$$

$$(5)\qquad w^t u = 0.$$

There are now $n+2$ equations in $n+2$ unknowns which can be solved to give the vector w in the direction of ℓ_2.

The additional step of the procedure is thus to find the line ℓ_2 for each group and observe the resulting histogram of the projected points on ℓ_2. In this manner the possibility of missing a group is reduced.

The termination of the elongated ellipsoid analysis completes the cluster analysis procedure. In summary the method is described in the following steps.

(1) Determine the optimal line using the non-linear programming method shown previously.

(2) Construct a histogram representing the distribution of the projected points.

(3) Examine the histogram and form a group for each peak.

(4) Repeat Steps 1, 2, and 3 for each maximum and its associated group until none will sub-divide.

(5) Analyze the combinations of adjacent groups.

(6) Examine each group to insure that it cannot be reasonably split into two groups.

It is to be noted that the implementation of the final step need not be confined to the end. For the most favorable result, when the points of a region are projected on their absolute maximum variance line, they should also be projected on their relative maximum variance line. Then, if a division of the region is made, it should be done in accordance with the histogram on that line ℓ_1 or ℓ_2 which most clearly divides the points.

G. INTERPRETATIONS

In light of the earlier discussion of bell-shaped histograms, some interesting observations concerning the various groups can now be made. As noted previously, the more closely the distribution of points in a cluster

approximates a hyperball, the more well defined and bell-like is the curve-fitted histogram representing the distribution of points when orthogonally projected on the optimal line. Thus, simply by observing the histogram for each group, it is possible to make comparisons between groups and to derive useful information about the probable shape of the hypervolume spanned by the points.

Probably the most influential factor in making a judgement will revolve around some within group variance study. In this case, the magnitude of the variance would not be a sufficient comparable statistic between groups; it could easily occur that group r_i and group r_j are equally dense, where density is points per unit volume, but group r_j is larger and contains twice as many points as group r_i. This results in group r_j having a characteristically larger variance, even though both groups may span a similarly shaped hypervolume.

One meaningful comparable statistic could be formed as follows: (1) calculate the standardized variance, i.e., the variance assuming the entire line segment for each group has length one and (2) form a ratio between the height of the mode to that variance, assuming each point in an interval contributes a unit of height to the curve. Observe that the smaller this ratio, the more elongated is the ellipse; in fact, in the limit as the ratio tends to zero, the point distribution approaches a straight line segment.

As a further observation it can be seen that the shape of the ellipsoid yields information concerning the original n specimen variates. Knowledge that a group is an elongated ellipsoid in conjunction with knowledge of its orientation in space is useful in eliciting information about the measured quantities. The process of ascertaining the dominant axis direction, or ellipsoid orientation, is accomplished in the following manner: as before, that line onto which the variance of the orthogonally projected points is a maximum can be determined. This will be that line which is approximately parallel to the longest axis of the ellipsoid and hence, the orientation is specified. Then two particularly interesting cases can occur: (1) the direction of the line may coincide closely with a coordinate axis, that is, the "cigar-shaped" distribution of points may have its longest axis approximately parallel to one of the axis of the coordinate system or (2) the distribution of points may be oriented such that its longest axis lies directionally near the forty-five degree (45^{o}) line between two coordinate axes.

Consider Figure 13, an illustration of the first case.

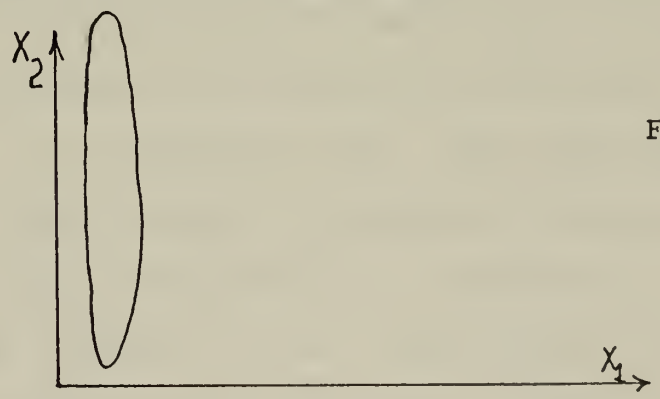


Figure 13

It should be realized that an elongated ellipsoid parallel to an axis indicates that the observations of the sundry characteristics are not varying comparably. That is, one dimension is exerting a relatively powerful force in the determination of point dispersion. Said differently, the observations of one characteristic are varying considerably while observations of the other fluctuate only slightly. Furthermore, an increase in one variate does not affect the other. Thus, for the n-dimensional case, the characteristic represented by the coordinate axis which is parallel to the major axis of the ellipsoid is independent of all other coordinate axes.

The second case, in which a different orientation of the ellipsoid is considered, will first be examined in two-dimensional space. Figure 14 will help to illuminate the discussion.

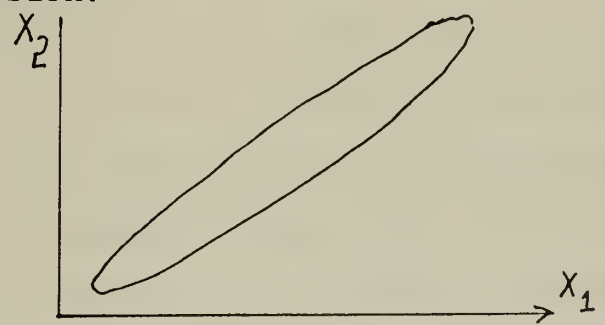


Figure 14

The elongated ellipsoid has its major axis running in a direction near parallel to the forty-five degree line between the x_1 and x_2 coordinate axes. The implication in this event is that there exists an interaction between the two measured specimen characteristics represented by x_1 and x_2, hence, the variates are not independent. As an example, suppose that blood pressure and age of humans were being

measured. Generally speaking, blood pressure rises with age and it can be expected that an increase in age would effect an increase in blood pressure. Thus, the expected distribution of points would be in the form of an elongated ellipse whose major axis is positively sloping and directionally close to the forty-five degree line between the representative coordinate axes. An elongated ellipse negatively sloping approximately 135^{o} between two axes indicates a complimentary trade-off between specimen variates; an increase in one variate is characteristically accompanied by a decrease in the other. Extending this concept to n-dimensional space, it can be seen that if the orthogonal projection of a group in the plane formed by any two dimensions results in a "cigar-shaped" distribution of points oriented directionally close to the forty-five degree line between the axes, then the two specimen variates represented by those axes are probably directly related. Said differently, if the unit vector in the direction of the line which maximizes the variance of the projected points of an elongated ellipsoid, as found using the non-linear programming method, contains two direction cosines which are approximately equal, then there may exist a dependence between the underlying variates. In addition, if it is desired to know more precisely the relation between two variates a linear regression analysis can be performed.

H. CAUTION

Although the procedure presented in this paper is capable of handling a wide variety of point dispersion patterns, there still exists many cases in which the techniques of this method are inadequate. For example, consider Figures 15a and 15b.

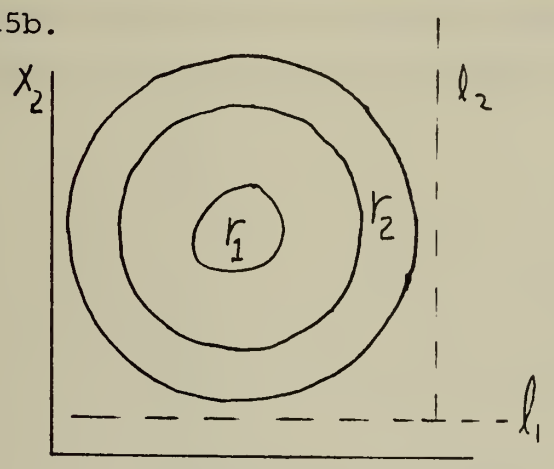


Figure 15a

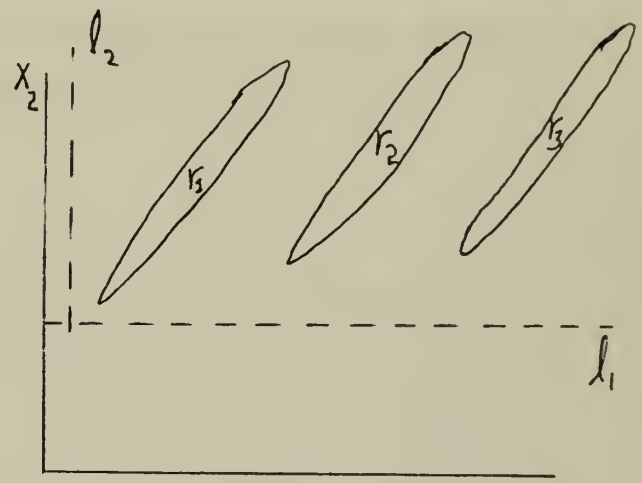


Figure 15b

In Figure 15a it is not possible to detect the smaller group completely surrounded by a "donut-shaped" group. The n-dimensional analog of this is a hyperball (solid sphere) contained in a hypersphere (hollow sphere). Again it can be seen that only one group would be detected. In Figure 15b ℓ_1 is the maximum variance line and ℓ_2 is perpendicular to ℓ_1. It is possible that there exists precisely that amount of overlap between the group projections on ℓ_1 to cause the histogram to be unimodal. In that case only one group is detected using ℓ_1. When line ℓ_2 is examined again only one group is detected. In either case A or B it is clear that it is not desirable to term the entire dispersion of points one cluster.

Many of the rules set forth in this work are designed to handle a general case. The user may wish to alter some of these to fit his specific data, his motives for using cluster analysis and his desired accuracy. For example, depending upon the specific criteria selected the two-dimensional point dispersion shown in Figure 16 might be termed as one group or as two by the method presented in this paper.

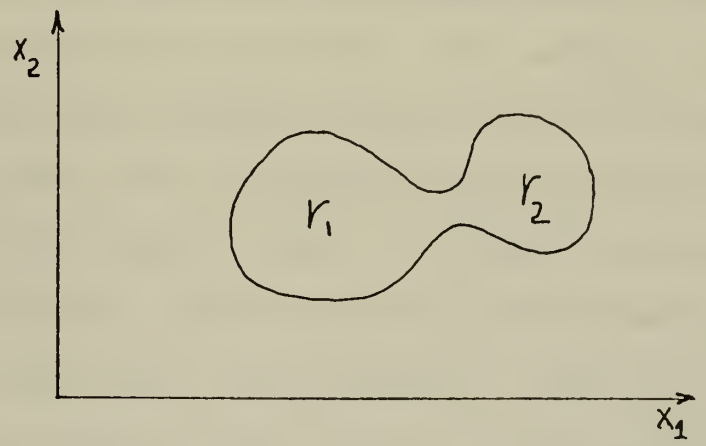


Figure 16

The reader may decide if the point dispersion is one or two groups by relaxing or tightening the constraint for making divisions when combining adjacent groups.

III. EXAMPLE

A. EXPLANATION OF THE DATA

To illustrate the operation of the analysis procedure, an example was chosen in which thirteen variates were measured on 254 specimens. The specimens are naval officers who attended the Naval Postgraduate School, Monterey, during the years 1964 through 1966. The variates are the results of several measurements derived from the Allport, Vernon, and Lindzey (AVL) and the Edwards Personal Preference Schedule (EPPS) psychological tests. Through the theory of psychology and personality, as set forth in Edward Spranger's Types of Men, the two above tests measure a total of twenty-one aspects of human nature, six in the first test and fifteen in the second. To avoid repetition, only thirteen of these variates are used in the cluster analysis example, six from AVL and seven from the EPPS.

The six scores from the AVL reflect the following aspects of human make-up: (1) theoretical, which pertains to the desire to pursue truth, (2) economic, or the interest in what is useful, (3) aesthetic, which involves the interest in form and harmony, (4) social, or the high esteem for love of people and unselfishness, (5) political, or lust for power and (6) religious, which involves the desire for unity or comprehension of the whole.

Those variates chosen from the EPPS measure: (1) achievement, or the ability to accomplish tasks requiring skill and

effort, (2) deference, which involves the acceptance or acknowledgement of the leadership of others, (3) affiliation, or loyalty to friends or groups, (4) intraception, or the ability to analyze the behavior of others, (5) dominance, or the facility to be a leader, (6) nurturance, which is giving of oneself to others who need kindness or understanding and (7) endurance, or the ability to see a job through to completion.

It is desirable to know if the 254 naval officers cluster as one high point density area or as several when the above thirteen variates are measured on each. If there do exist several groups, the characteristics of each could be analyzed and perhaps yield valuable information to the Navy. For example, if most of the officers in one group were those who voluntarily resigned, the AVL and the EPPS could then be used as a predictor to aid the Navy in determining beforehand those officers who would probably not make the service a career. Such a result would indeed have great usefulness within the Navy. On the other hand, if points representing the officers cluster into only one group, the indication is that the AVL and EPPS would not be successful as predictors for officer career patterns.

B. PROCEDURE

The first step in the cluster analysis method is to project the points on their maximum variance line. The Fortran IV program included at the end of this paper first determines the set of non-linear equations to be solved,

then determines their solution, which is in the form of the unit vector in the direction of the maximum variance line, and ultimately projects all points on that line. The only inputs to the program necessary are the coordinates of all the points and the number of dimensions. Using this program it was found that the initial set of fourteen simultaneous non-linear equations to be solved when considering all 254 points are:

$$AU - 2 \lambda U = 0$$

$$\sum_{i=1}^{13} u_i^2 - 1 = 0$$

where

$$U = (u_1, u_2, \ldots, u_{13})$$

and

A is the thirteen by thirteen array of numbers defined on the following page.

A =

33.3	-3.9	-6.7	-1.2	5.9	-12.0	3.8	10.5	-.7	47.2	-6.7	6.1	-15.7
3.9	22.6	2.6	3.1	-1.9	.8	-.2	-2.5	-2.8	3.0	-2.1	-3.2	7.5
-6.7	2.6	31.4	2.7	-4.3	17.4	-8.7	-3.0	-1.2	-6.1	4.3	-2.9	7.6
-1.2	3.1	2.7	47.7	-.2	.6	2.6	-3.2	-6.7	-5.2	1.7	.9	12.9
5.9	-1.9	-4.3	-.2	36.1	-6.7	.1	6.6	9.3	-7.1	-1.8	-13.1	-18.5
-12.0	.8	17.4	.6	-6.7	36.1	5.2	-6.0	-5.3	-10.6	6.4	-6.8	21.5
3.8	-.2	-8.7	2.6	.1	-5.2	49.7	14.2	4.5	-4.6	-7.8	-11.7	5.9
10.5	2.5	3.0	-3.2	6.6	-6.0	14.2	67.3	5.5	-.2	-23.0	-7.1	-40.8
-.7	-2.8	-1.2	-6.7	9.3	-5.3	4.5	5.5	99.9	-64.2	-38.7	38.6	-38.3
7.2	3.0	-6.1	-5.2	-7.1	-10.6	-9.6	-.2	-64.2	165.5	3.5	-55.9	-48.7
-6.7	-2.1	4.3	1.7	-1.8	6.4	-7.8	-23.0	-38.7	3.5	94.2	-42.2	6.1
6.1	-3.2	-2.9	.9	13.1	-6.8	-11.7	-7.1	38.6	55.9	-42.2	101.8	-30.6
-15.7	7.5	7.6	12.9	-18.5	21.5	5.9	-40.8	-38.3	-48.7	6.0	-30.6	154.2

Their solution in terms of u_1 through u_{13} is

$$u = (-.32, .72, -.23, -.06, -.08, -.18, -.10, -.07, -.26, -.26, -.16, -.17, -.23)$$

This vector, u, is the unit vector, u, in the direction of the maximum variance line. The 254 points were then projected on the line whose direction is given by the above unit vector. As can be seen by the first histogram in the appendix, the curve-fitted histogram of the projected points is unimodal. Therefore, the points were projected on the relative maximum variance line, which is perpendicular to the absolute maximum variance line. The curve-fitted histogram as shown in the appendix contains two local maxima and a division can be made in accordance with the division rule requiring the local minima to be no greater than eight-tenths the height of the lesser adjacent peaks. Thus, n-dimensional space has been partitioned into two mutually exclusive and collectively exhaustive regions r_1, containing 142 points and r_2, containing the remaining 112 points. The next two curves indicate that the region r_1 would not subdivide further, either when its points were projected on the absolute maximum variance line or the relative maximum variance line. When region r_2 was considered it was found that a division was possible when the points were projected on their absolute maximum variance line. Region r_2 was, therefore, partitioned into r_{2a} and r_{2b},

with fifty-two points allocated to the former and sixty to the latter.

As shown in the following diagrams, it was not possible to make any further division of any of the regions. Furthermore, region r_{2a} was found to be adjacent to r_{2b}. The two regions were combined into one, r_2, sinze the zero slope, positive second derivative point on the curve-fitted histogram had value approximately eighty percent the height of the lesser peak. Similarly, r_1 and r_2 were combined. According to this method, then, the points are one cluster. It is interesting to note that this result is in agreement with several different statistical techniques applied to this same data. These methods include various forms of regression analysis and a discriminate analysis technique developed at The University of California at Berkeley and presented by J. W. Dixon, (Ref. 2).

C. CONCLUSION

The conclusion which must be reached in accordance with this cluster analysis on the given data is that the AVL and EPPS were not shown to be valid for distinguishing among naval officers with respect to the thirteen criteria mentioned. It should be noted, however, that the data used was very narrow-based; that is, it represented only a small portion of the officers in the Navy, namely a fraction of those attending graduate school. If a more representative sample could be obtained, it is indeed feasible that different results would be obtained.

IV. APPENDIX A :

HISTOGRAM ANALYSIS

This appendix contains ten histograms and their curve-fitted approximations. On each diagram is the unit vector specifying the line onto which the points were projected. In those cases where a division was made the division line is shown.

Following the histograms is a computer program which yields sufficient data to facilitate a histogram analysis. The program first determines the set of non-linear equations to be solved for the unit vector, it secondly determines their solution, and thirdly projects the points onto the line specified by the unit vector. Thus, for each histogram shown, a set of non-linear equations was found and solved for the unit vector and the points were subsequently projected on the specified line by the computer program.

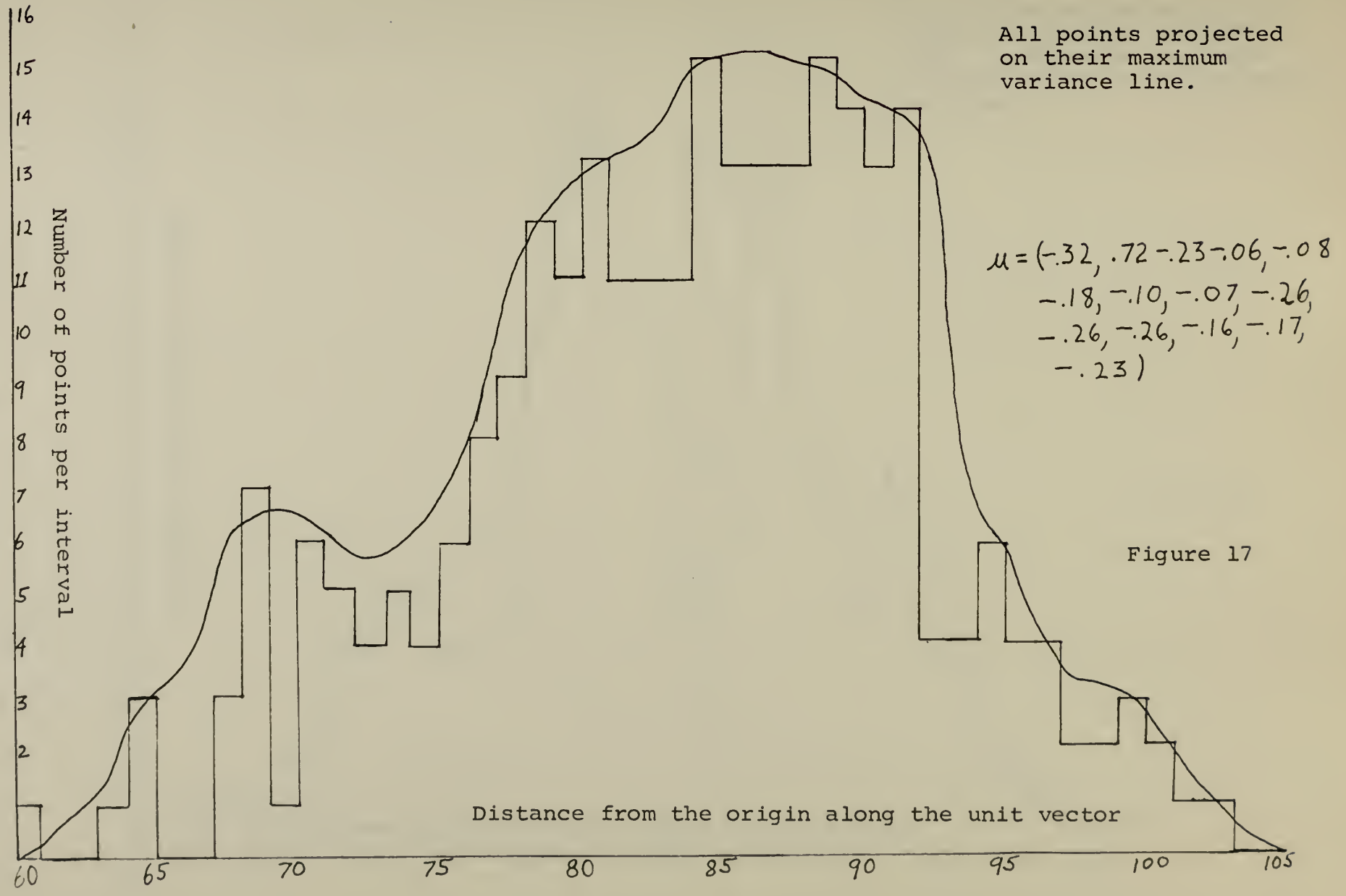
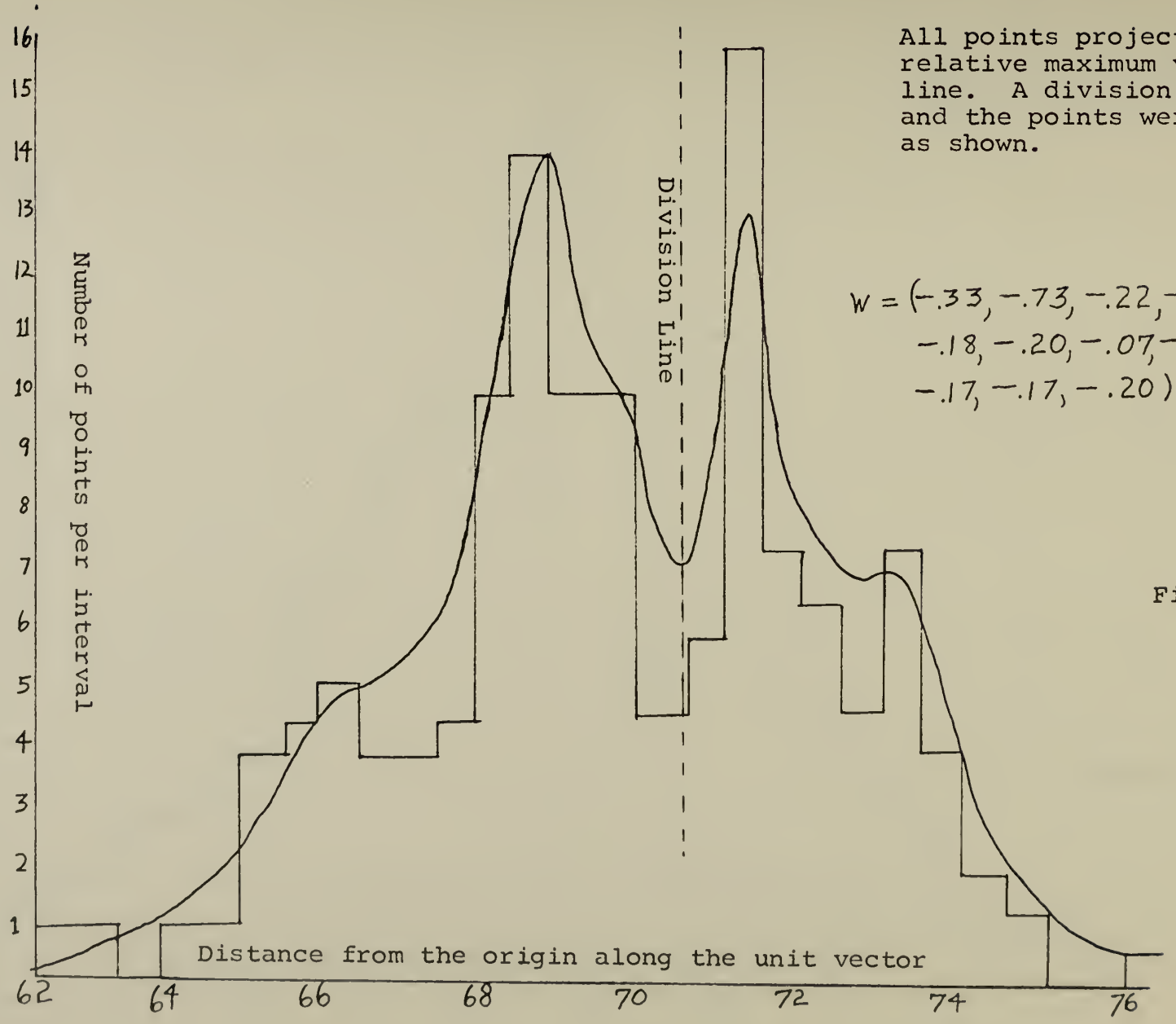

All points projected on their maximum variance line.
μ = (-.32, .72 -.23 -.06, -.08 -.18, -.10, -.07, -.26, -.26, -.26, -.16, -.17, -.23)
Number of points per interval
Distance from the origin along the unit vector
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
60
65
70
75
80
85
90
95
100
105

Figure 17



All points projected on the relative maximum variance line. A division was made and the points were separated as shown.

$W = (-.33, -.73, -.22, -.05, -.06$
$-.18, -.20, -.07, -.26, -.25$
$-.17, -.17, -.20)$

Figure 18

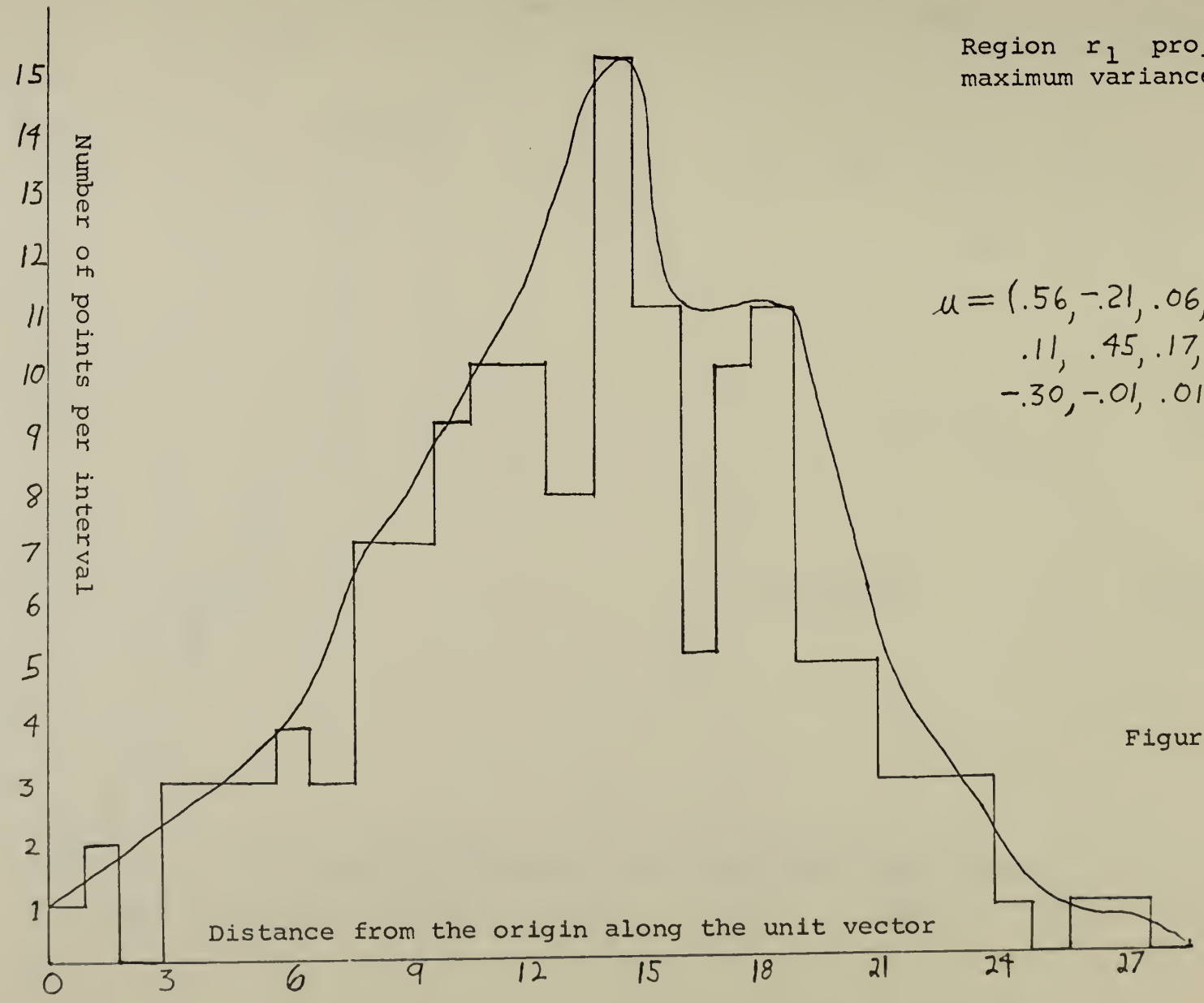

Region r_1 projected on its maximum variance line.
$u = (.56, -.21, .06, -.33, -.37$
$.11, .45, .17, .13, .12,$
$-.30, -.01, .01)$
Number of points per interval
Distance from the origin along the unit vector
0
3
6
9
12
15
18
21
24
27
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 19

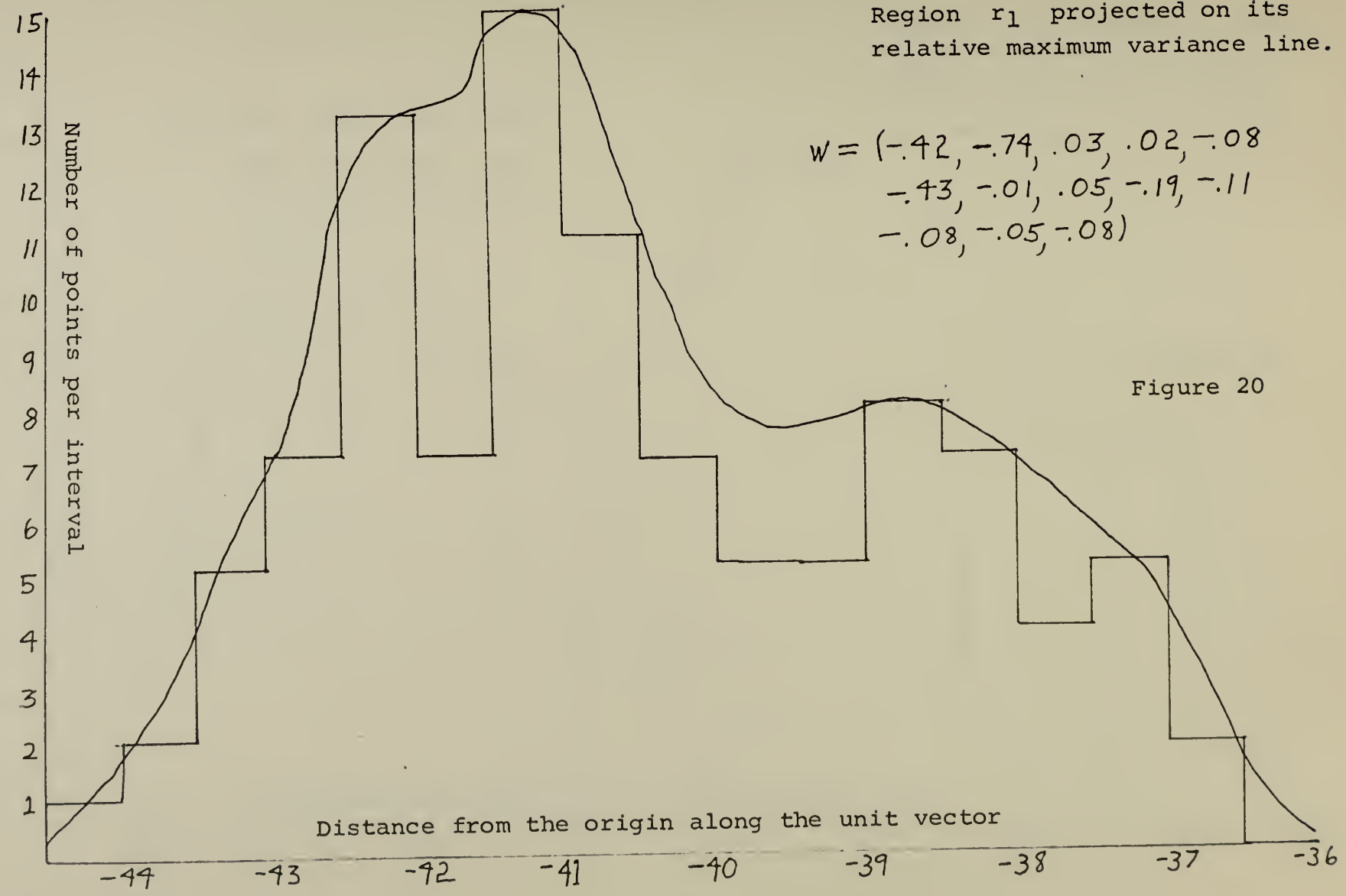

Region r_1 projected on its relative maximum variance line.
$w = (-.42, -.74, .03, .02, -.08$
$-.43, -.01, .05, -.19, -.11$
$-.08, -.05, -.08)$
Number of points per interval
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Distance from the origin along the unit vector
-44
-43
-42
-41
-40
-39
-38
-37
-36

Figure 20

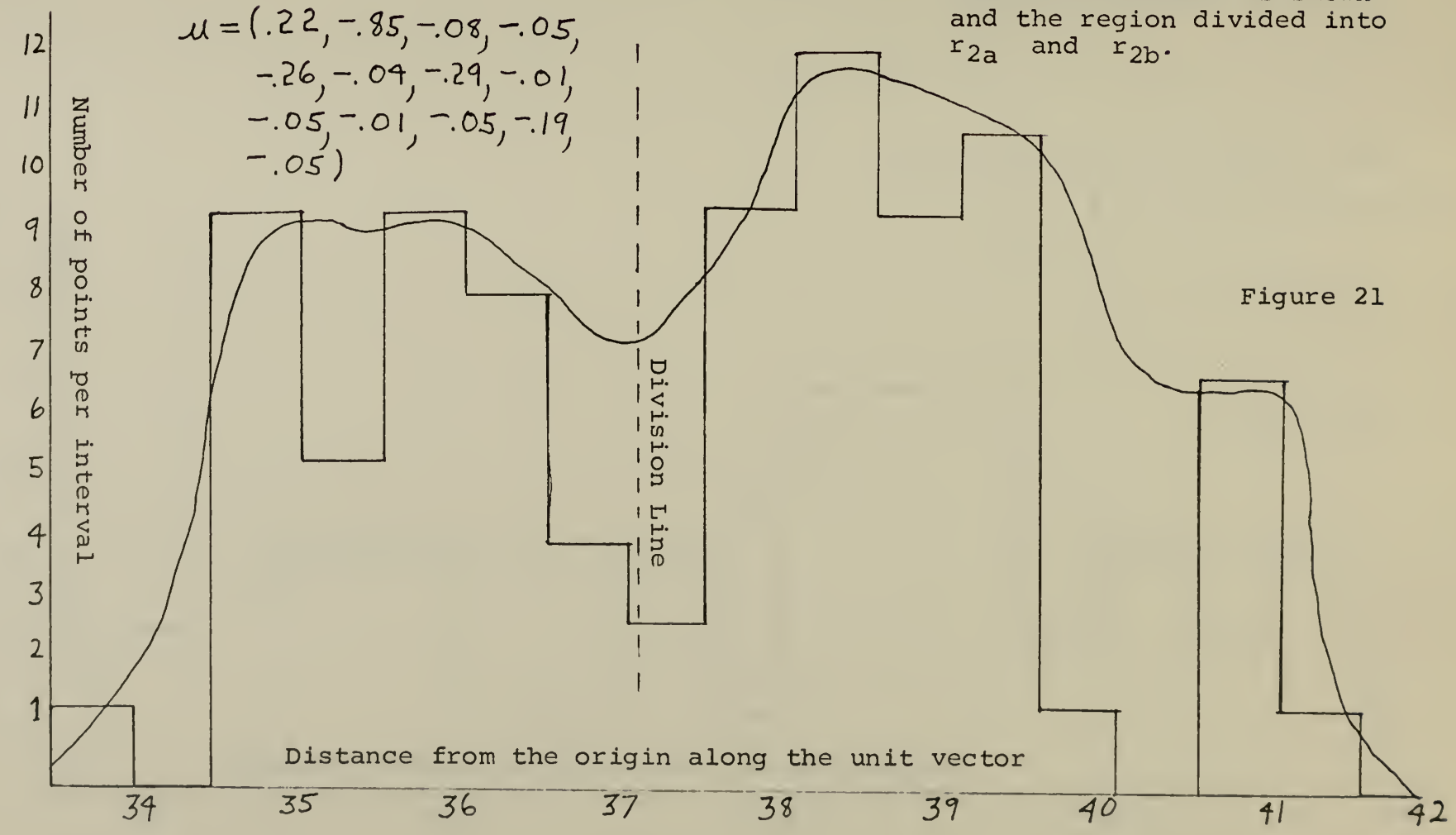

Region r_2 projected on its maximum variance line. A division was made as shown and the region divided into r_{2a} and r_{2b}.
$\mu = (.22, -.85, -.08, -.05, -.26, -.04, -.29, -.01, -.05, -.01, -.05, -.19, -.05)$
Number of points per interval
Division Line
Distance from the origin along the unit vector
12
11
10
9
8
7
6
5
4
3
2
1
34
35
36
37
38
39
40
41
42

Figure 21

Region r_2 projected on its relative maximum variance line.

$w = (-.25, -.85, -.19, .07, .01, -.02$
$-.15, .13, -.13, -.13, -.15, -.18$
$-.17)$

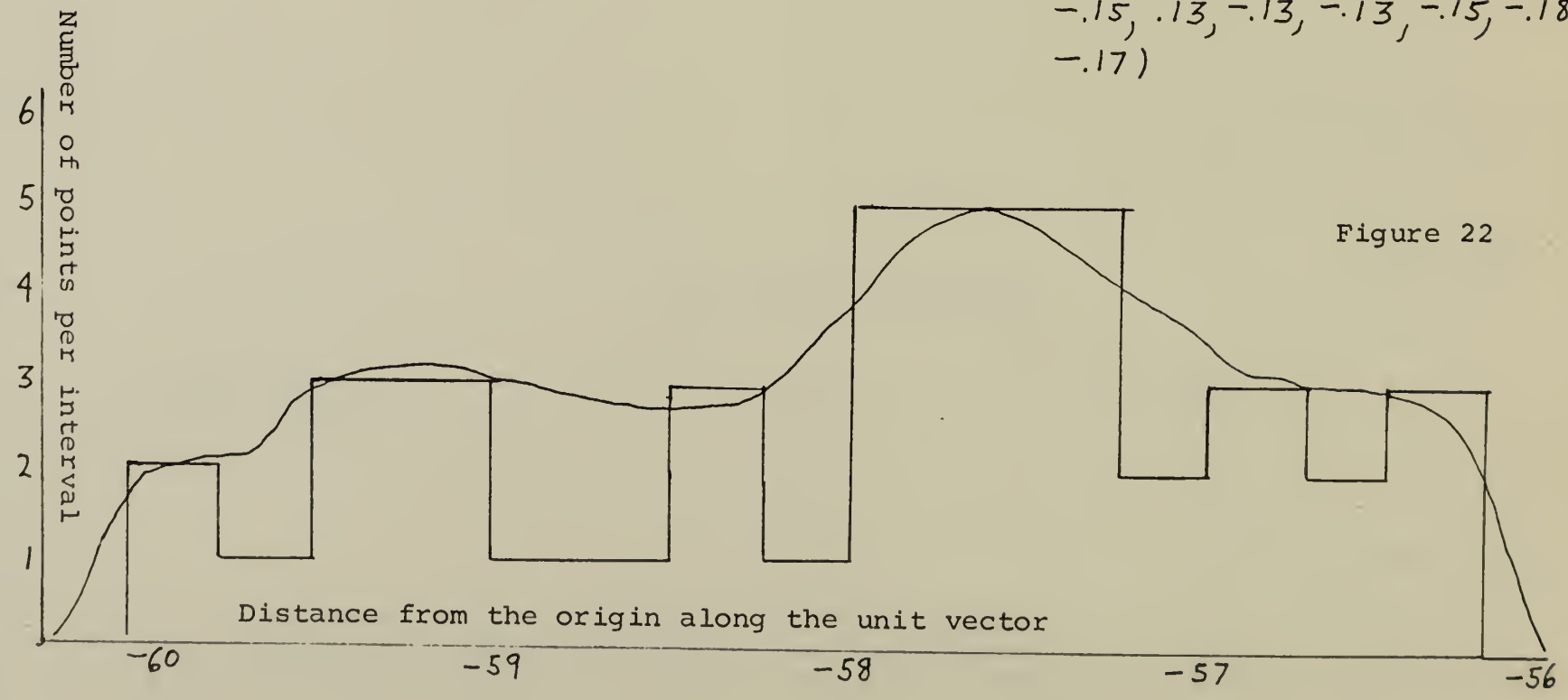


Figure 22

Region r_{2a} projected on its maximum variance line.

$$\mu = (-.02, -.43, .01, .13, .60, .41, .36, -.32, .03, .02, -.11, -.05, -.03)$$

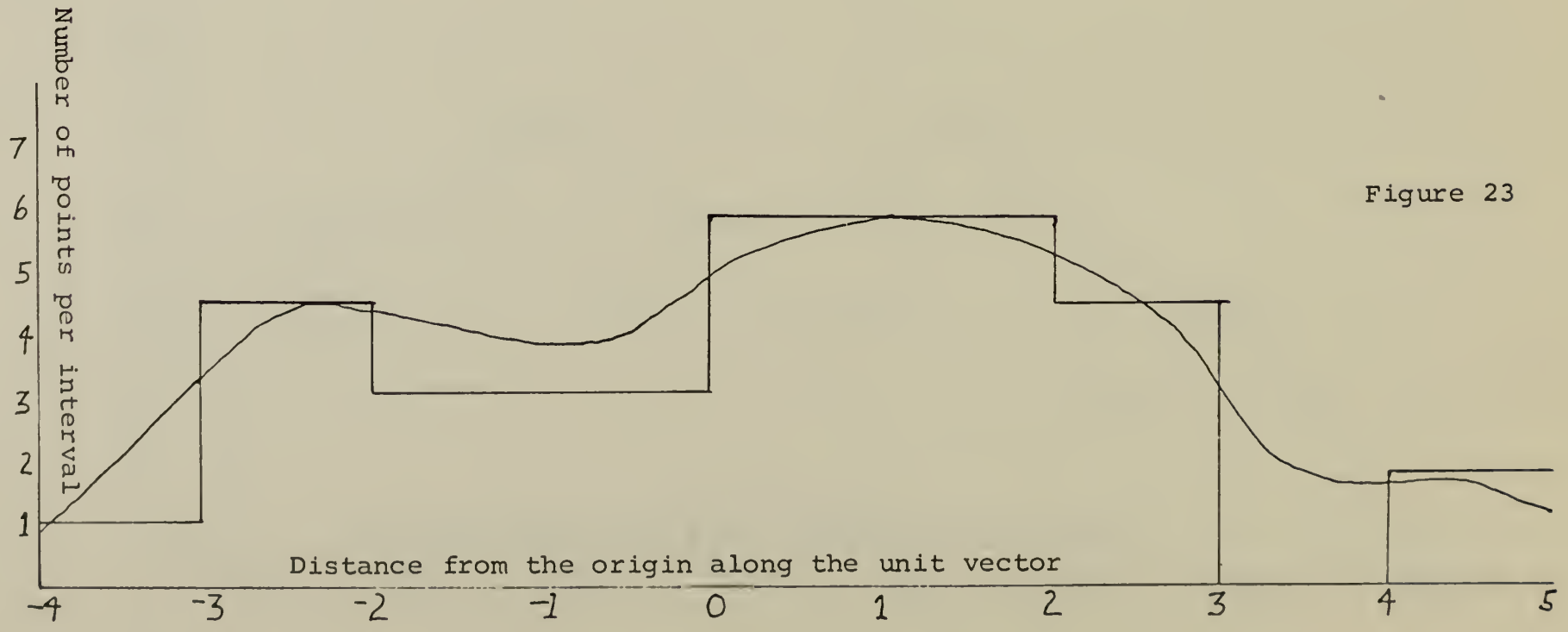


Figure 23

Region r_{2a} projected on its relative maximum variance line.

$w = (-.17, -.56, -.31, .24, .35, .30$
$.10, -.30, -.13, -.16, -.25,$
$-.19, -.18)$

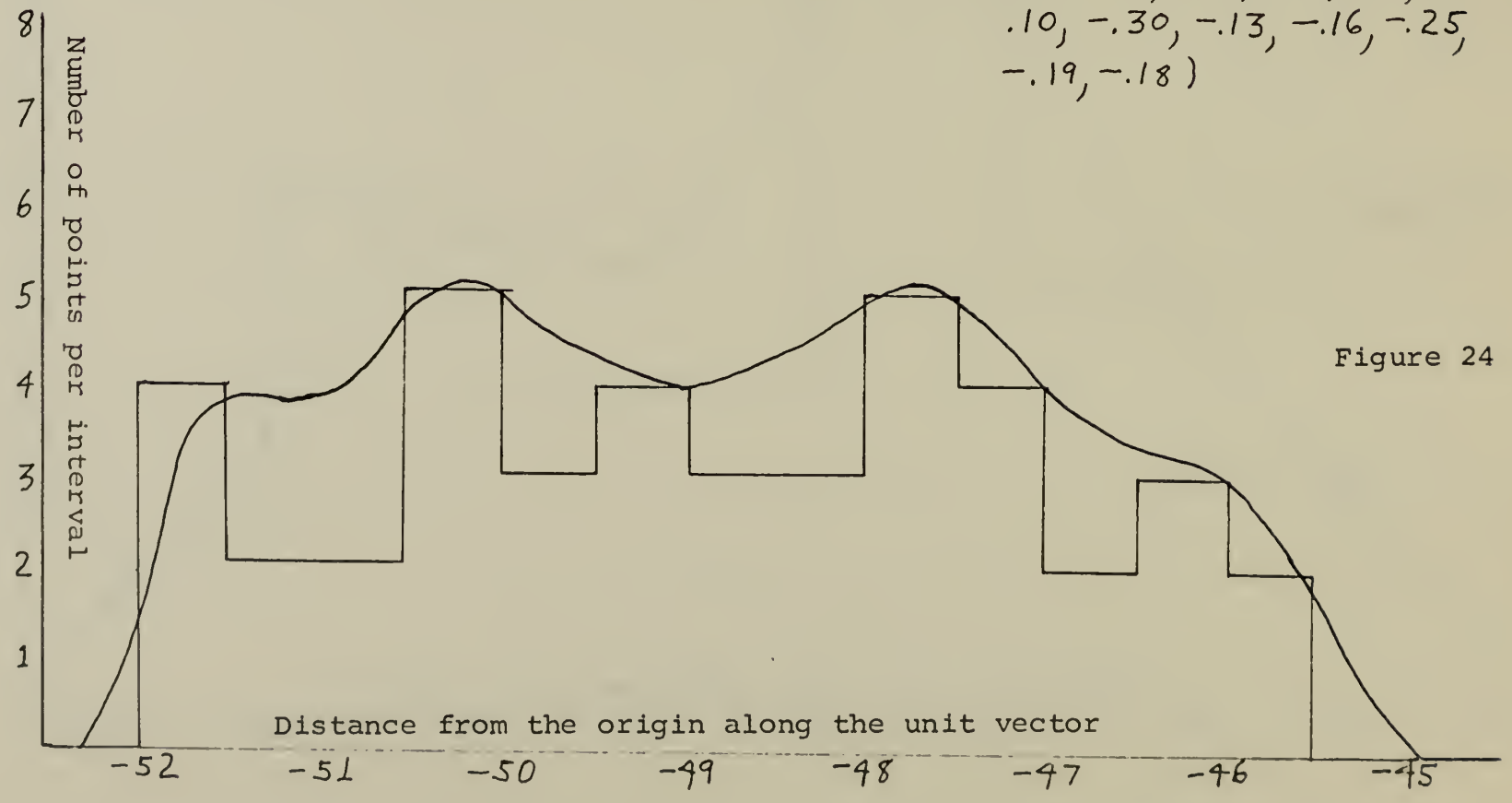


Figure 24

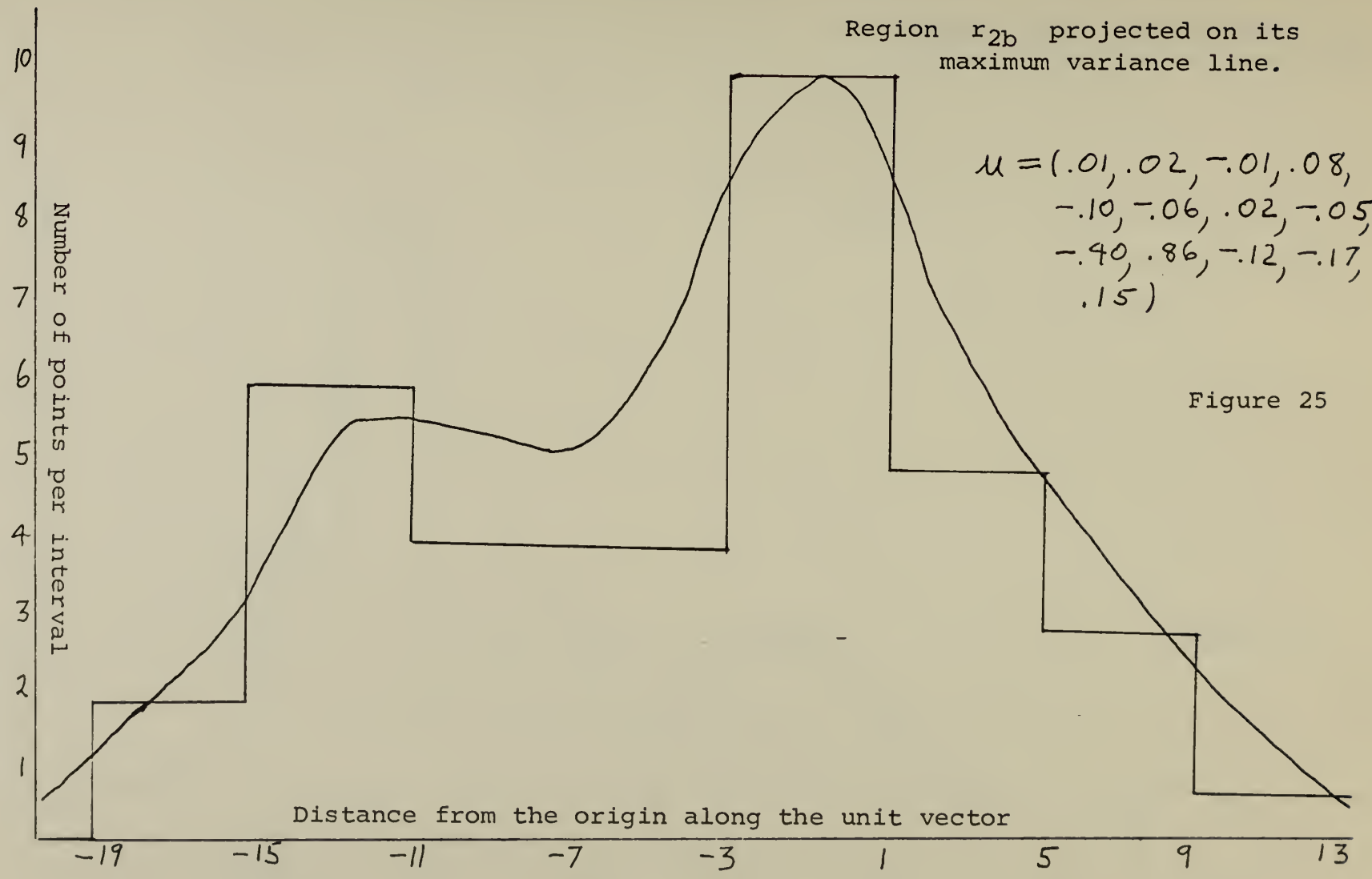

Region r_{2b} projected on its maximum variance line.
$\mu = (.01, .02, -.01, .08, -.10, -.06, .02, -.05, -.40, .86, -.12, -.17, .15)$
Number of points per interval
Distance from the origin along the unit vector
10
9
8
7
6
5
4
3
2
1
-19
-15
-11
-7
-3
1
5
9
13

Figure 25

Region r_{2b} projected on its relative maximum variance line.

$w = (-.11, -.30, .64, -.05, .02, -.69, -.13, -.04, .05, -.03, -.09, -.17, .03)$

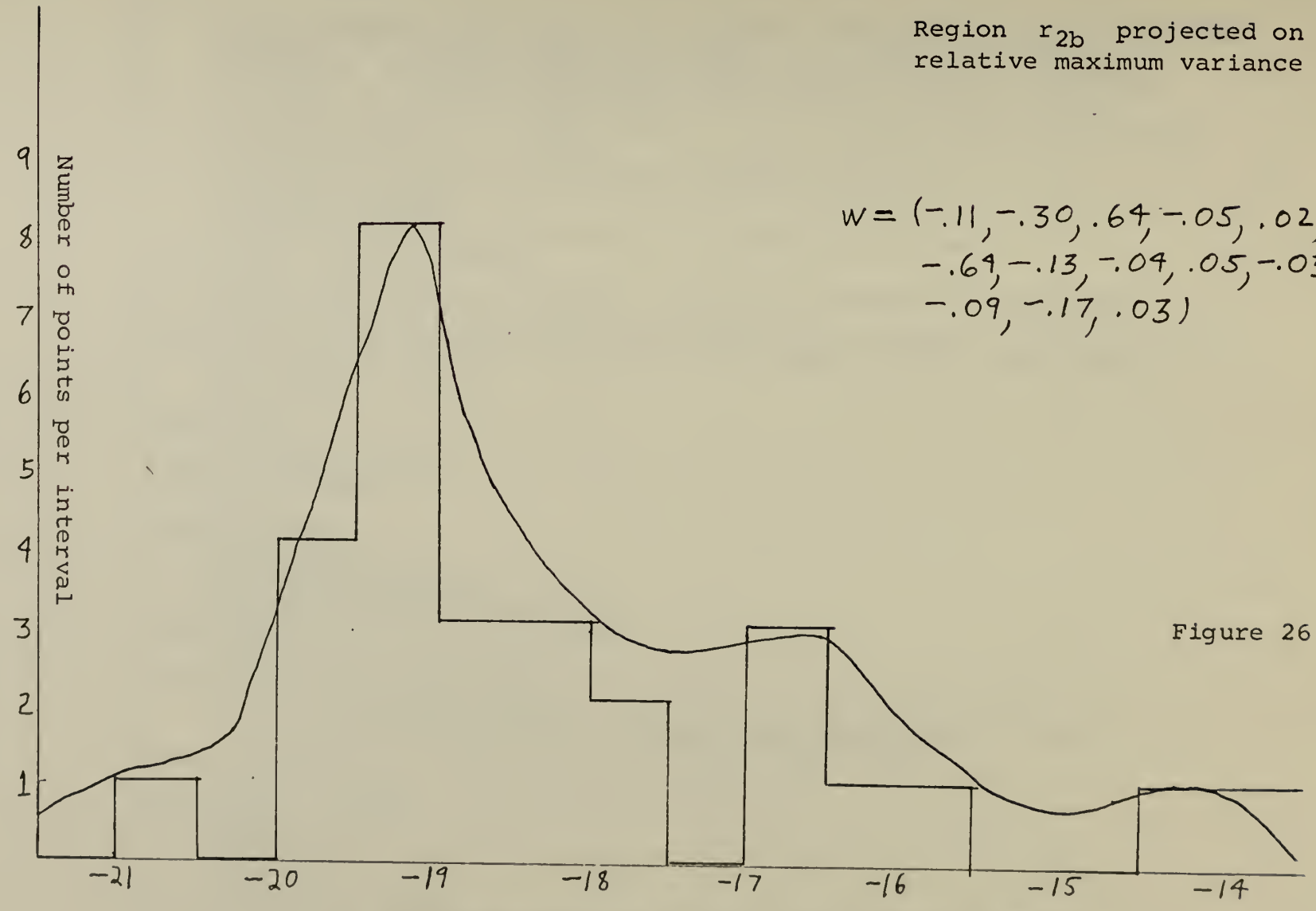


Figure 26

```
C     THIS PROGRAM IS DESIGNED TO DO SEVERAL PHASES OF THE COMPUTATIONAL
CC    WORK IN THE CLUSTER ANALYSIS METHOD.  IT FIRST DETERMINES THE
      SYSTEM OF N SIMULTANEOUS NON-LINEAR EQUATIONS TO BE SOLVED.
CC    SECONDLY, IT SOLVES THESE EQUATIONS FOR THE OPTIMAL (MAXIMUM
      VARIANCE OR RELATIVE MAXIMUM VARIANCE) LINE AND ULTIMATELY
C     PROJECTS POINTS IN A UNIT VECTOR IN THE DIRECTION OF THIS LINE.
      DIMENSION UU(13),FF(259)
      DIMENSION X(28)
      DIMENSION C(259,13),B(259,13),U(13,13),D(259,13)
      EXTERNAL EVALUT
      INTEGER POINTER(30,30),CDE(30,30),ISUB(30),U,MM,NN
      COMMON POINTER,CDE,ISUB
      READ (5,2001) MM,NN
 2001 FORMAT (2I4)
C     MM IS THE NUMBER OF POINTS.  NN IS THE NUMBER OF DIMENSIONS.
      READ (5,2002) ((C(I,J),J=1,NN),I=1,MM)
 2002 FORMAT (18X,2F20.0,6X,2F20.0,2X,1F20.0,2X,1F20.0/20X,6F20.0)
C     C'S ARE THE POINTS.  C(I,J), J=1,...,NN, ARE THE COORDINATES
C     OF THE ITH POINT.
      NN=NN+1
      K=1
 2003 SUM=0.0
      DO 2010 I=1,MM
 2010 SUM=SUM+C(I,K)
      SUM=SUM/MM
      J=1
 2004 CONTINUE
      U(K,J)=0.0
      DO 2020 I=1,MM
      B(I,K)=C(I,K)-SUM
      D(I,K)=2.0*B(I,K)/(MM-1.0)
 2020 U(K,J)=D(I,K)*C(I,J)+U(K,J)
      J=J+1
      IF(J-NN)2004,2004,2005
 2005 K=K+1
      IF(K-NN)2003,2003,2006
C     AT THIS POINT ALL OF THE U VALUES HAVE BEEN FOUND.  U(K,J)
C     IS THE MULTIPLIER OF X(J) IN THE KTH EQUATION.
 2006 CONTINUE
      WRITE(6,2015)  ((U(K,J),J=1,NN),K=1,NN)
 2015 FORMAT (F9.4)
      N=27
      MAXIT=1000
      NUMSIG=4
      IPRINT=1
      DO 4000 II=1,NN
 4000 X(II)=26
      X(14)=5.0
```

```
      X(28)=5.0
      NNNN=NN+2
      DO 4050 II=NNNN,N
 4050 X(II)=1.0
      CALL NLNSYS(N,MAXIT,NUMSIG,ISING,IPRINT,EVALUT,X)
      IF (ISING.EQ.0) WRITE (6,2000)
      GO TO 3001
C     AT THIS POINT, THE SET OF EQUATIONS HAS BEEN SOLVED AND POINTS WILL NOW BE
C     AT THIS POINT, THE SET OF EQUATIONS HAS BEEN SOLVED AND POINTS
C     WILL NOW BE PROJECTED ON THE LINE.
 2000 FORMAT(1H1,55HSYSTEM CREATED A SINGULARITY NEAR THE FOLLOWING X VA
     1LUE)
 3001 CONTINUE
      SUM=0.0
      DO 3022 I=1,NN
 3022 SUM=SUM+X(I)**2
      SUM1=SQRT(SUM)
      DO 3006 J=1,NN
 3006 UU(J)=X(J)/SUM1
      WRITE (6,1111)
 1111FORMAT (20H  THE UNIT VECTOR IS)
      WRITE (6,1000) (UU(J),J=1,NN)
 1000 FORMAT (F8.4)
C     UU(J) IS THE JTH COMPONENT OF THE UNIT VECTOR.
      DO 3005 I=1,MM
      SUM=0.0
      DO 3010 J=1,NN
 3010 SUM=SUM+C(I,J)*UU(J)
 3005 FF(I)=SUM
      DO 3023 I=1,MM
      WRITE (6,3024) I
 3024 FORMAT (I4)
 3015 FORMAT (F12.4)
 3023 WRITE (6,3015) FF(I)
C     FF(I) IS A CONSTANT WHICH WHEN MULTIPLIED TIMES THE UNIT VECTOR
C     GIVES THE COORDINATE OF THE ITH POINT WHEN PROJECTED OF THE LINE.
      STOP
      END


      SUBROUTINE EVALUT(X,K,F)
C     IF IT IS DESIRED TO PROJECT POINTS ON THE MAXIMUM VARIANCE LINE,
C     THE EQUATIONS ARE AS SHOWN BELOW.  IT WILL BE NOTED THAT THEIR
C     FORM IS IDENTICAL TO THAT SHOWN IN THE THESIS BODY.  THERE ARE
C     2NN+1 EQUATIONS, WHERE NN IS THE NUMBER OF DIMENSIONS.  THE FIRST
C     NN OF THESE ARE THE FIRST DERIVATIVES OF THE LAGRANGIAN, THE
C     SECOND NN ARE THE SECOND DERIVATIVES OF THE LAGRANGIAN AND
C     THE FINAL EQUATION IS THE CONSTRAINT EQUATION..  NOTICE THAT
```

```
C      IT IS SPECIFIED THAT ALL SECOND DERIVATIVES BE LESS THAN ZERO.
C      IN THE EVENT THAT IT IS DESIRED TO PROJECT POINTS ON THE RELATIVE
C      MAXIMUM VARIANCE LINE, IT IS FIRST NECESSARY TO FIND THE UNIT
C      VECTOR IN THE DIRECTION OF THE MAXIMUM VARIANCE LINE.  THEN,
C      A TER MUST BE ADDED TO EACH OF THE FIRST NN EQUATIONS OF THE FORM -
C      -UU(I)*X(29) FOR THE ITH EQUATION, I=1,...,NN, WHERE UU(I)
C      IS THE ITH COMPONENT OF THE UNIT VECTOR IN THE DIRECTION OF THE
C      MAXIMUM VARIANCE LINE.  IN ADDITION, ANOTHER EQUATION MUST BE
C      ADDED TO SATISFY THE CONSTRAINT THAT THE RELATIVE MAXIMUM
C      VARIANCE LINE BE PERPENDICULAR TO THE ABSOLUTE MAXIMUM VARIANCE
C      LINE.  THIS EQUATION WILL BE OF THE FORM, SUM OVER I OF
C      +UU(I)*X(I), I=1,...,NN
       DIMENSION X(29)
       DIMENSION U(13,13)
       INTEGER PONTER
       COMMON PONTER(30,30),COE(30,30),ISUB(30),U,MM,NN
       IF(K-1) 1001,1001,1002
 1001 F=(U(1,1)-2.0*X(14))*X(1)+U(1,2)*X(2)+U(1,3)*X(3)+U(1,4)*X(4)+U(1,
     15)*X(5)+U(1,6)*X(6)+U(1,7)*X(7)+U(1,8)*X(8)+U(1,9)*X(9)+U(1,10)*X(
     210)+U(1,11)*X(11)+U(1,12)*X(12)+U(1,13)*X(13)
       GO TO 1004
 1002 IF (K-2) 1003,1003,1005
 1003 F=U(2,1)*X(1)+(U(2,2)-2.0*X(14))*X(2)+U(2,3)*X(3)+U(2,4)*X(4)+U(2,
     15)*X(5)+U(2,6)*X(6)+U(2,7)*X(7)+U(2,8)*X(8)+U(2,9)*X(9)+U(2,10)*X(
     210)+U(2,11)*X(11)+U(2,12)*X(12)+U(2,13)*X(13)
       GO TO 1004
 1005 IF(K-3) 1006,1006,1007
 1006 F=U(3,1)*X(1)+U(3,2)*X(2)+(U(3,3)-2.0*X(14))*X(3)+U(3,4)*X(4)+U(3,
     15)*X(5)+U(3,6)*X(6)+U(3,7)*X(7)+U(3,8)*X(8)+U(3,9)*X(9)+U(3,10)*X(
     210)+U(3,11)*X(11)+U(3,12)*X(12)+U(3,13)*X(13)
       GO TO 1004
 1007 IF(K-4) 1008,1008,1009
 1008 F=U(4,1)*X(1)+U(4,2)*X(2)+U(4,3)*X(3)+(U(4,4)-2.0*X(14))*X(4)+U(4,
     15)*X(5)+U(4,6)*X(6)+U(4,7)*X(7)+U(4,8)*X(8)+U(4,9)*X(9)+U(4,10)*X(
     210)+U(4,11)*X(11)+U(4,12)*X(12)+U(4,13)*X(13)
       GO TO 1004
 1009 IF (K-5) 1010,1010,1011
 1010 F=U(5,1)*X(1)+U(5,2)*X(2)+U(5,3)*X(3)+U(5,4)*X(4)+(U(5,5)-2.0*X(14
     1))*X(5)+U(5,6)*X(6)+U(5,7)*X(7)+U(5,8)*X(8)+U(5,9)*X(9)+U(5,10)*X(
     210)+U(5,11)*X(11)+U(5,12)*X(12)+U(5,13)*X(13)
       GO TO 1004
 1011 IF(K-6) 1012,1012,1013
 1012 F=U(6,1)*X(1)+U(6,2)*X(2)+U(6,3)*X(3)+U(6,4)*X(4)+U(6,5)*X(5)+(U(6
     1,6)-2.0*X(14))*X(6)+U(6,7)*X(7)+U(6,8)*X(8)+U(6,9)*X(9)+U(6,10)*X(
     210)+U(6,11)*X(11)+U(6,12)*X(12)+U(6,13)*X(13)
       GO TO 1004
 1013 IF(K-7) 1014,1014,1015
 1014 F=U(7,1)*X(1)+U(7,2)*X(2)+U(7,3)*X(3)+U(7,4)*X(4)+U(7,5)*X(5)+U(7,
```

```
     16)*X(6)+(U(7,7)-2.0*X(14))*X(7)+U(7,8)*X(8)+U(7,9)*X(9)+U(7,10)*X(
     210)+U(7,11)*X(11)+U(7,12)*X(12)+U(7,13)*X(13)
      GO TO 1004
1015 IF(K-8) 1016,1016,1017
1016 F=U(8,1)*X(1)+U(8,2)*X(2)+U(8,3)*X(3)+U(8,4)*X(4)+U(8,5)*X(5)+U(8,
     16)*X(6)+U(8,7)*X(7)+(U(8,8)-2.0*X(14))*X(8)+U(8,9)*X(9)+U(8,10)*X(
     210)+U(8,11)*X(11)+U(8,12)*X(12)+U(8,13)*X(13)
      GO TO 1004
1017 IF (K-9) 1018,1018,1019
1018 F=U(9,1)*X(1)+U(9,2)*X(2)+U(9,3)*X(3)+U(9,4)*X(4)+U(9,5)*X(5)+U(9,
     16)*X(6)+U(9,7)*X(7)+U(9,8)*X(8)+(U(9,9)-2.0*X(14))*X(9)+U(9,10)*X(
     210)+U(9,11)*X(11)+U(9,12)*X(12)+U(9,13)*X(13)
      GO TO 1004
1019 IF (K-10) 1020,1020,1021
1020 F=U(10,1)*X(1)+U(10,2)*X(2)+U(10,3)*X(3)+U(10,4)*X(4)+U(10,5)*X(5)
     1+U(10,6)*X(6)+U(10,7)*X(7)+U(10,8)*X(8)+U(10,9)*X(9)+(U(10,10)-2.0
     2*X(14))*X(10)+U(10,11)*X(11)+U(10,12)*X(12)+U(10,13)*X(13)
      GO TO 1004
1021 IF(K-11) 1022,1022,1023
1022 F=U(11,1)*X(1)+U(11,2)*X(2)+U(11,3)*X(3)+U(11,4)*X(4)+U(11,5)*X(5)
     1+U(11,6)*X(6)+U(11,7)*X(7)+U(11,8)*X(8)+U(11,9)*X(9)+U(11,10)*X(10
     2)+(U(11,11)-2.0*X(14))*X(11)+U(11,12)*X(12)+U(11,13)*X(13)
      GO TO 1004
1023 IF(K-12) 1024,1024,1025
1024 F=U(12,1)*X(1)+U(12,2)*X(2)+U(12,3)*X(3)+U(12,4)*X(4)+U(12,5)*X(5)
     1+U(12,6)*X(6)+U(12,7)*X(7)+U(12,8)*X(8)+U(12,9)*X(9)+U(12,10)*X(10
     2)+U(12,11)*X(11)+(U(12,12)-2.0*X(14))*X(12)+U(12,13)*X(13)
      GO TO 1004
1025 IF(K-13) 1026,1026,1027
1026 F=U(13,1)*X(1)+U(13,2)*X(2)+U(13,3)*X(3)+U(13,4)*X(4)+U(13,5)*X(5)
     1+U(13,6)*X(6)+U(13,7)*X(7)+U(13,8)*X(8)+U(13,9)*X(9)+U(13,10)*X(10
     2)+U(13,11)*X(11)+U(13,12)*X(12)+(U(13,13)-2.0*X(14))*X(13)
      GO TO 1004
1027 IF (K-14) 1028,1028,1053
1028 F=U(1,1)-2.0*X(14)+X(15)**2
      GO TO 1004
1053 IF (K-15) 1030,1030,1031
1030 F=U(2,2)-2.0*X(14)+X(16)**2
      GO TO 1004
1031 IF(K-16) 1032,1032,1033
1032 F=U(3,3)-2.0*X(14)+X(17)**2
      GO TO 1004
1033 IF(K-17) 1034,1034,1035
1034 F=U(4,4)-2.0*X(14)+X(18)**2
      GO TO 1004
1035 IF(K-18) 1036,1036,1037
1036 F=U(5,5)-2.0*X(14)+X(19)**2
      GO TO 1004
```

```
 1037 IF(K-19) 1038,1038,1039
 1038 F=U(6,6)-2.0*X(14)+X(20)**2
      GO TO 1004
 1039 IF (K-20) 1040,1040,1041
 1040 F=U(7,7)-2.0*X(14)+X(21)**2
      GO TO 1004
 1041 IF (K-21) 1042,1042,1043
 1042 F=U(8,8)-2.0*X(14)+X(22)**2
      GO TO 1004
 1043 IF(K-22) 1044,1044,1045
 1044 F=U(9,9)-2.0*X(14)+X(23)**2
      GO TO 1004
 1045 IF(K-23) 1046,1046,1047
 1046 F=U(10,10)-2.0*X(14)+X(24)**2
      GO TO 1004
 1047 IF(K-24) 1048,1048,1049
 1048 F=U(11,11)-2.0*X(14)+X(25)**2
      GO TO 1004
 1049 IF(K-25) 1050,1050,1051
 1050 F=U(12,12)-2.0*X(14)+X(26)**2
      GO TO 1004
 1051 IF(K-26) 1052,1052,1029
 1052 F=U(13,13)-2.0*X(14)+X(27)**2
      GO TO 1004
 1029 F=X(1)**2+X(2)**2+X(3)**2+X(4)**2+X(5)**2+X(6)**2+X(7)**2+X(8)**2+
     1X(9)**2+X(10)**2+X(11)**2+X(12)**2+X(13)**2-1.0
 1004 RETURN
      END

      SUBROUTINE NLNSYS(N,MAXIT,NUMSIG,ISING,IPRINT,EVALUT,X)                NLNS1430
C                                                                            NLNS0000
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC NLNS0010
CCCCCCC                                                                      NLNS0020
CCCCCCC   SUBROUTINE NLNSYS                                                  NLNS0030
CCCCCCC                                                                      NLNS0040
CCCCCCC   IDENTIFICATION                                                     NLNS0050
CCCCCCC      TITLE       - NLNSYS   (SOLUTION OF SIMULTANEOUS NON-LINEAR     NLNS0060
CCCCCCC                               EQUATIONS).                            NLNS0070
CCCCCCC      ID          - C4-NPG-NLNSYS (FORTRAN IV G).                     NLNS0080
CCCCCCC      CATEGORY    -MATHEMATICAL SUBROUTINE                            NLNS0090
CCCCCCC      PROGRAMMER  - D. D. PURCELL                                      NLNS0100
CCCCCCC      DATE        - FEB. 1969                                         NLNS0110
CCCCCCC                                                                      NLNS0120
CCCCCCC   PURPOSE                                                            NLNS0130
CCCCCCC      TO SOLVE A SYSTEM OF N NON-LINEAR EQUATIONS.  THESE MUST BE     NLNS0140
CCCCCCC      SUPPLIED BY THE USER IN A SUBROUTINE REFERENCED BY THE MAIN-     NLNS0150
C            LINE CALLING ROUTINE.                                           NLNS0160
```

```
C                                                                       NLNS0170
C     USAGE                                                             NLNS0180
C        CALL NLNSYS(N,MAXIT,NUMSIG,ISING,IPRINT,EVALUT,X)              NLNS0190
C                                                                       NLNS0200
C     DESCRIPTION OF PARAMETERS                                         NLNS0210
C        N      - THE NUMBER OF EQUATIONS.                             NLNS0220
C        MAXIT  - THE ITERATION LIMIT.                                 NLNS0230
C        NUMSIG - THE NUMBER OF SIGNIFICANT PLACES REQUIRED IN THE     NLNS0240
C                 SOLUTION.  IT SHOULD BE LESS THAN OR EQUAL TO 7,     NLNS0250
C                 WHICH IS THE LIMIT WITH SINGLE PRECISION.            NLNS0260
C        ISING  - A RETURN PARAMETER SUCH THAT:                        NLNS0270
C                 1. ISING IS 1 IF THE STARTING GUESS TOGETHER WITH    NLNS0280
C                    THE SYSTEM OF EQUATIONS DOES NOT GENERATE A       NLNS0290
C                    SINGULARITY DURING THE PROCESS.                   NLNS0300
C                 2. ISING IS 2 IF THE STARTING GUESS TOGETHER WITH    NLNS0310
C                    THE SYSTEM OF EQUATIONS DOES GENERATE A SING-     NLNS0320
C                    ULARITY DURING THE PROCESS.                       NLNS0330
C        IPRINT - USER SETS THIS TO ZERO IF HE WANTS THE SUCCESSIVE    NLNS0340
C                 ITERATIVE APPROXIMATIONS TO BE PRINTED (ON UNIT 6).  NLNS0350
C                 ANYTHING ELSE WILL OVERIDE THIS PRINT-OUT.           NLNS0360
C        EVALUT - THE SUBROUTINE CREATED BY THE USER WHICH SPECIFIES    NLNS0370
C                 THE NON-LINEAR SYSTEM AND RETURNS THE VALUE OF A     NLNS0380
C                 PARTICULAR MEMBER FOR SOME X VECTOR. ITS FORM AND    NLNS0390
C                 PARAMETERS ARE SUBSEQUENTLY EXPLAINED.               NLNS0400
C                 SUBROUTINE EVALUT(X,F,K)    WHERE X IS A SOLUTION    NLNS0410
C                                               VECTOR ESTIMATE        NLNS0420
C                                               GENERATED BY           NLNS0430
C                                               NLNSYS, AND            NLNS0440
C                                             F IS THE VALUE OF        NLNS0450
C                                               THE                    NLNS0460
C                                             K'TH FUNCTION FOR        NLNS0470
C                                               THAT X VALUE.          NLNS0480
C        X      - THE STARTING GUESS VECTOR AT CALLING TIME AND THE    NLNS0490
C                 SOLUTION WHEN CONTPOL IS RETURNED TO THE MAINLINE.   NLNS0500
C     EXAMPLE                                                           NLNS0510
C        DIMENSION X(2)                                                 NLNS0520
C        EXTERNAL FUNCS                                                 NLNS0530
C        X(1)=1.0                                                       NLNS0540
C        X(2)=4.0                                                       NLNS0550
C        CALL NLNSYS(2,10,4,ISING,X,FUNCS)                              NLNS0560
C        IF (ISING.EQ.0) WRITE (6,2000)                                 NLNS0570
C   2000 FORMAT(1H1,55HSYSTEM CREATED A SINGULARITY NEAR THE FOLLOWIN   NLNS0580
C       1G X VALUE)                                                     NLNS0590
C        WRITE (6,1000) X(1),X(2)                                       NLNS0600
C   1000 FORMAT(1H1,4HX1 =,F10.7,10X,4HX2 =,F10.7)                      NLNS0610
C        END                                                            NLNS0620
C                                                                       NLNS0630
C        SUBROUTINE FUNCS(X,K,F)                                        NLNS0640
```

```
C
C     THE PARAMETERS ARE RESPECTIVELY - SOLUTION VECTOR GUESS,
C        NUMBER OF FUNCTION TO BE RETURNED, AND THE RETURN PAR-
C        AMETER REPRESENTING THE VALUE OF THIS FUNCTION FOR THIS
C        APPROXIMATION.
      DIMENSION X(2)
      IF (K.EQ.1) GO TO 1
      GO TO 2
    1 F=2.718281828*(.9204225528*(EXP(2.*X(1)-1.)-1.)+X(2))/
     1   3.141592652-2.*X(1)
      GO TO 3
    2 F=.5*SIN(X(1)*X(2))-X(2)/12.5663706-X(1)/2.
    3 RETURN
      END

C SPACE REQUIRED
C    NLNSYS PLUS BAKSUB REQUIRE 2400 BYTES AND THE SUPPORT SYS-
C    TEMS ROUTINES PLUS BUFFERS REQUIRE ABOUT 43,000 BYTES.

C CAUTIONS TO USER
C    THE STARTING GUESS MUST BE WITHIN THE REGION NECESSARY FOR
C    CONVERGENCE OR THE PROCESS WILL 'BLOW UP'.  THIS REGION HAS
C    NO EASY DEFINITION AND IT MAY BE NECESSARY TO TRY MORE THAN
C    ONE STARTING GUESS.  ALSO, THE USER SUBROUTINE MUST BE TYPED
C    'EXTERNAL' IN THE CALLING PROGRAM.
C EQUIPMENT CONFIGURATION
C    IBM 360
C SUBROUTINE AND FUNCTION SUBPROGRAMS REQUIRED
C    EVALUT (USER WRITTEN AND USER NAMED)
C    BAKSUB (SUPPLIED WITH NLNSYS)
C    STANDARD SYSTEMS SUBPROGRAMS.

C REMARKS
C    THIS PROCEDURE SOLVES A SYSTEM OF N SIMULTANEOUS NON-
C    LINEAR EQUATIONS.  THE METHOD IS ROUGHLY QUADRATICALLY CON-
C    VERGENT AND REQUIRES ONLY ((N**2/2)+(3*N/2)) FUNCTION
C    EVALUATIONS PER ITERATIVE STEP AS COMPARED WITH (N**2+N)
C    EVALUATIONS FOR NEWTON'S METHOD.  THIS RESULTS IN A SAVINGS
C    OF COMPUTATIONAL EFFORT FOR SUFFICIENTLY COMPLICATED FUNC-
C    TIONS.  A DETAILED DESCRIPTION OF THE GENERAL METHOD AND
C    PROOF OF CONVERGENCE ARE INCLUDED IN (2).  BASICALLY THE
C    TECHNIQUE CONSISTS IN EXPANDING THE FIRST EQUATION IN A TAY-
C    LOR SERIES ABOUT THE STARTING GUESS, RETAINING ONLY LINEAR
C    TERMS, EQUATING TO ZERO AND SOLVING FOR ONE VARIABLE IN
C    X(K), AS A LINEAR COMBINATION OF THE REMAINING N-1 VARI-
C    ABLES.  IN THE SECOND EQUATION, X(K) IS ELIMINATED BY RE-
C    PLACING IT WITH ITS LINEAR REPRESENTATION FOUND ABOVE, AND
C    AGAIN THE PROCESS OF EXPANDING THROUGH LINEAR TERMS OF
C    EQUATING TO ZERO AND SOLVING FOR ONE VARIABLE IN TERMS OF
```

```
C           THE NOW REMAINING N-2 VARIABLES IS PERFORMED.  ONE CONTINUES NLNS1130
C           IN THIS FASHION,ELIMINATING ONE VARIABLE PER EQUATION,       NLNS1140
C           UNTIL FOR THE N'TH EQUATION, WE ARE LEFT WITH ONE EQUATION   NLNS1150
C           IN ONE UNKNOWN.  A SINGLE NEWTON STEP IS NOW PERFORMED, FOL- NLNS1160
C           LOWED BY BACK-SUBSTITUTION IN THE TRIANGULARIZED LINEAR SYS- NLNS1170
C           TEM GENERATED FOR THE X(I)'S.  A PIVOTING EFFECT IS ACHIEVED NLNS1180
C           BY CHOOSING FOR ELIMINATION AT ANY STEP THAT VARIABLE HAVING NLNS1190
C           A PARTIAL DERIVATIVE OF LARGEST ABSOLUTE VALUE.  THE PIV-    NLNS1200
C           OTING IS DONE WITHOUT PHYSICAL INTERCHANGE OF ROWS OR COL-   NLNS1210
C           UMNS.                                                        NLNS1220
C                                                                        NLNS1230
C       REFERENCES                                                       NLNS1240
C           1. BROWN, K. M. "ALGORITHM 316 SOLUTION OF SIMULTANEOUS NON- NLNS1250
C              LINEAR EQUATIONS (C5)", COLLECTED ALGORITHMS FROM CACM,   NLNS1260
C              JULY 1967, 316-P 1- 0 AND I- 0.                           NLNS1270
C                           THE REMARKS IMMEDIATELY ABOVE ARE RE-        NLNS1280
C                           PRODUCED FROM THIS WRITE-UP.                 NLNS1290
C           2. BROWN, K. M. A QUADRATICALLY CONVERGENT METHOD FOR        NLNS1300
C              SOLVING SIMULTANEOUS NON-LINEAR EQUATIONS.  DOCTORAL      NLNS1310
C              THESIS, DEPT. COMPUTER SCIENCES, PURDUE U., LAFAYETTE,    NLNS1320
C              IND., AUG., 1966.                                         NLNS1330
C           3. BROWN, K. M., AND CONTE, S. D.  THE SOLUTION OF SIM-      NLNS1340
C              ULTANEOUS NONLINEAR EQUATIONS.  PROC. ACM 22ND NAT.       NLNS1350
C              CONF., PP 111-114.                                        NLNS1360
C                                                                        NLNS1370
C     ..................................................................NLNS1380
C                                                                        NLNS1390
C                                                                        NLNS1400
C                                                                        NLNS1410
C                                                                        NLNS1420
C                                                                        NLNS1430
      DIMENSION X(1)                                                     NLNS1440
      DIMENSION U(13,13)
      INTEGER CONVRG,TALLY,PONTER                                        NLNS1460
      COMMON PONTER(30,30),COE(30,30),ISUB(30),U,MM,NN
      EXTERNAL EVALUT                                                    NLNS1480
      DIMENSION TEMP(30),PART(30)
C                                                                        NLNS1500
      CONVRG=1                                                           NLNS1510
C        INITIALIZE CONVERGENCE AND TOLERANCE DETERMINERS.               NLNS1520
      ISING=1                                                            NLNS1530
      RELCON=10.**(-NUMSIG)                                              NLNS1540
C        ITERATION LOOP.                                                 NLNS1550
      DO 55 M=1,MAXIT                                                    NLNS1560
C        PROGRAM WILL PRINT OUT SUCCESSIVE APPROXIMATIONS OF X IF        NLNS1570
C        IPRINT IS 0.                                                    NLNS1580
      IF (IPRINT.EQ.0) WRITE (6,1000) M,(X(I),I=1,N)                     NLNS1590
 1000 FORMAT(1X,12HON ITERATION,I4,17HTHE X ESTIMATE IS,5E15.7/7E15.7/   NLNS1600
```

```
     1       7F15.7///)                                                        NLNS1610
C       INITIALIZE FIRST ROW OF ARRAY WHICH ENABLES INDEXING TO PIVOT         NLNS1620
C       THROUGH THE VARIABLES WITHOUT INTERCHANGING ROWS OR COLUMNS.          NLNS1630
        DO 5 J=1,N                                                            NLNS1640
      5 PONTER(1,J)=J                                                         NLNS1650
C       ONE K FOR EACH EQUATION.                                              NLNS1660
        DO 40 K=1,N                                                           NLNS1670
C       IF K IS GREATER THAN 1 BACK-SUBSTITUTION IS NECESSARY.                NLNS1680
        IF (K.GT.1) CALL BAKSUB(K,N,X)                                        NLNS1690
        CALL EVALUT(X,K,F)                                                    NLNS1700
        FACTOR=.001                                                           NLNS1710
      7 TALLY=0                                                               NLNS1720
        DO 10 I=K,N                                                           NLNS1730
        ITEMP=PONTER(K,I)                                                     NLNS1740
        HOLD=X(ITEMP)                                                         NLNS1750
C       GET INCREMENT TO OBTAIN ITEMP'TH PARTIAL.                             NLNS1760
        H=FACTOR*HOLD                                                         NLNS1770
        IF (H.EQ.C.) H=.001                                                   NLNS1780
        X(ITEMP)=HOLD+H                                                       NLNS1790
C       IF K IS GREATER THAN 1, A NEW X(ITEMP) VALUE WILL AFFECT X'S          NLNS1800
C       WHICH ARE EXPRESSED IN TERMS OF IT AS A RESULT OF PREVIOUS            NLNS1810
C       EQUATION STEPS.                                                       NLNS1820
        IF (K.GT.1) CALL BAKSUB(K,N,X)                                        NLNS1830
        CALL EVALUT(X,K,FPLUS)                                                NLNS1840
C       GET ITEMP'TH PARTIAL.                                                 NLNS1850
        PART(ITEMP)=(FPLUS-F)/H                                               NLNS1860
        X(ITEMP)=HOLD                                                         NLNS1870
C       IF PARTIAL IS TOO SMALL, INCREASE TALLY.                              NLNS1880
        IF(ABS(PART(ITEMP)).NE.C.) GO TO 80
     85 TALLY=TALLY+1
        GO TO 10
     80 IF(ABS(F/PART(ITEMP)).GT.1.0E20) GO TO 85
     10 CONTINUE                                                              NLNS1910
        IF (TALLY.LE.(N-K)) GO TO 15                                          NLNS1920
        FACTOR=FACTOR*10.0                                                    NLNS1930
C       IF SURFACE IS TOO FLAT, THE SINGULARITY INDICATOR IS SET TO           NLNS1940
C       ZERO AND RETURN IS EXECUTED.                                          NLNS1950
        IF (FACTOR.GT..5) GO TO 65                                            NLNS1960
        GO TO 7                                                               NLNS1970
     15 IF(K.LT.N) GO TO 20                                                   NLNS1980
C       IF LAST PARTIAL IS ZERO, A SINGULARITY IS INDICATED AND A             NLNS1990
C       RETURN EXECUTED.                                                      NLNS2000
        IF (ABS(PART(ITEMP)).EQ.0.) GO TO 65                                  NLNS2010
        COF(K,N+1)=0                                                          NLNS2020
        KMAX=ITEMP                                                            NLNS2030
        GO TO 40                                                              NLNS2040
     20 KMAX=PONTER(K,K)                                                      NLNS2050
        DERMAX=ABS(PART(KMAX))                                                NLNS2060
```

```
      KPLUS=K+1                                                          NLNS2070
C        GET INDEX FOR LARGEST PARTIAL IN K'TH EQUATION.                 NLNS2080
      DO 30 I=KPLUS,N                                                    NLNS2090
      JSUB=PONTER(K,I)                                                   NLNS2100
      TEST=ABS(PART(JSUB))                                               NLNS2110
      IF (TEST.LT.DERMAX) GO TO 25                                       NLNS2120
      DERMAX=TEST                                                        NLNS2130
C        DEFINE PIVOT TO SWIVEL ABOUT THE VARIABLE WITH MAXIMUM PARTIAL  NLNS2140
C        WHEN WE GET TO THE NEXT EQUATION.                               NLNS2150
      PONTER(KPLUS,I)=KMAX                                               NLNS2160
C        IF THIS PARTIAL IS BIGGER, WE HAVE A NEW MAXIMUM.               NLNS2170
      KMAX=JSUB                                                          NLNS2180
      GO TO 30                                                           NLNS2190
   25 PONTER(KPLUS,I)=JSUB                                               NLNS2200
   30 CONTINUE                                                           NLNS2210
C        IF THAT PARTIAL IS 0, INDICATE A SINGULARITY AND RETURN.        NLNS2220
      IF (ABS(PART(KMAX)).EQ.0) GO TO 65                                 NLNS2230
      ISUB(K)=KMAX                                                       NLNS2240
      COE(K,N+1)=0                                                       NLNS2250
      DO 35 J=KPLUS,N                                                    NLNS2260
      JSUB=PONTER(KPLUS,J)                                               NLNS2270
C        SAVE THESE CONSTANTS FOR FUTURE USE.                            NLNS2280
      COE(K,JSUB)=-PART(JSUB)/PART(KMAX)                                 NLNS2290
C        GET PART OF EXPRESSION FOR THE NEW X(KMAX) VALUE                NLNS2300
   35 COE(K,N+1)=COE(K,N+1)+PART(JSUB)*X(JSUB)                           NLNS2310
   40 COE(K,N+1)=(COE(K,N+1)-F)/PART(KMAX)+X(KMAX)                       NLNS2320
C        IF N IS 1, WE HAVE OUR SOLUTION IN THE NEXT STEP WITHOUT ANY    NLNS2330
C        BACK-SUBSTITUTING.                                              NLNS2340
      X(KMAX)=COE(N,N+1)                                                 NLNS2350
C        FOR N GREATER THAN 1, WE PERFORM A FINAL BACK-SUBSTITUTION TO   NLNS2360
C        GET OUR NEW X-VECTOR.                                           NLNS2370
      IF (N.GT.1) CALL BAKSUB(N,N,X)                                     NLNS2380
      IF (M.EQ.1) GO TO 50                                               NLNS2390
      DO 43 I=1,N                                                        NLNS2400
C        TEST FOR CONVERGENCE.                                           NLNS2410
      IF (ABS((TEMP(I)-X(I))/X(I)).GT.RELCON) GO TO 45                   NLNS2420
   43 CONTINUE                                                           NLNS2430
      CONVRG=CONVRG+1                                                    NLNS2440
C        IF IT CONVERGES, RETURN WITH LAST VECTOR.                       NLNS2450
      IF (CONVRG.GE.3) GO TO 60                                          NLNS2460
      GO TO 50                                                           NLNS2470
   45 CONVRG=1                                                           NLNS2480
C        SAVE CURRENT X-VECTOR FOR TESTING WITH NEXT X-VECTOR.           NLNS2490
   50 DO 55 I=1,N                                                        NLNS2500
   55 TEMP(I)=X(I)                                                       NLNS2510
C        IF M IS THE ITERATION LIMIT, RETURN.                            NLNS2520
      GO TO 70                                                           NLNS2530
   60 MAXIT=M                                                            NLNS2540
```

```
      GO TO 70                                                          NLNS2550
   65 ISING=0                                                           NLNS2560
   70 RETURN                                                            NLNS2570
      END                                                               NLNS2580
C                                                                       BKSB2590

      SUBROUTINE BAKSUB(K,N,X)                                          BKSB2600
C        THIS SUBROUTINE BACK-SUBSTITUTES OR UPDATES VARIABLES WHICH ARBKSB2610
C        FUNCTIONS OF CURRENT X ENTRY VALUES.                           BKSB2620
      DIMENSION U(13,13)
      DIMENSION X(1)                                                    BKSB2630
      INTEGER PONTER                                                    BKSB2640
      COMMON PONTER(30,30),COE(30,30),ISUB(30),U,MM,NN
C                                                                       BKSB2660
      DO 10 KMM=2,K                                                     BKSB2670
      KM=K+2-KMM                                                        BKSB2680
      KMAX=ISUB(KM-1)                                                   BKSB2690
      X(KMAX)=0.                                                        BKSB2700
      DO 5 J=KM,N                                                       BKSB2710
      JSUB=PONTER(KM,J)                                                 BKSB2720
C        SEE (3) FOR THE EXPRESSION FOR X(KMAX).                        BKSB2730
    5 X(KMAX)=X(KMAX)+COE(KM-1,JSUB)*X(JSUB)                            BKSB2740
   10 X(KMAX)=X(KMAX)+COE(KM-1,N+1)                                     BKSB2750
      RETURN                                                            BKSB2760
      END                                                               BKSB2770
//GO.FT06F001 DD SYSOUT=A,DCB=BLKSIZE=133
//GO.SYSUDUMP  DD  SYSOUT=A
//GO.SYSIN  DD *
 254  13
  40205455          16131509130316071813061427172309
  4020                544926244740
  40216470          21131616150609021816061115212413
  4021                555454274010
  40265371          24141517121414082109041811181413
  4026                524427225144
  40334568          16061413141409072406122321131811
  4033                474230294944
  40504147          17102205160906142510111621171114
  4050                475137314034
  40594559          23111709110816032115121623132012
  4059                495434334129
  40625774          25121806120610231104081823191511
  4062                494038263948
  40694247          19101514180909072308082219281614
  4069                475733264829
  40705675          19152221201007062112011619041613
  4070                505034234637
```

(Sample data for 9 points)

VI. REFERENCES

1. Bonner, R. E., "On Some Clustering Techniques," IBM Journal of Research Development, v. 8, p. 22, 1964.

2. Dixon, J. W., Biomedical Computer Programs, Berkeley, pp. 196-206, 1968.

3. Dupraw, E. J., "Non-linnean Taxonomy," Nature, v. 202, pp. 849-852, May 30, 1964.

4. Edwards, A. W. F. and Cavalli-Sforza, "A Method for Cluster Analysis," Biometrics, v. 21, p. 62, 1964.

5. Gilmore, J. S. L., "The Development of Taxonomic Theory Since 1851," Nature, v. 168, pp. 400-402, 1951.

6. Gilmore, J. S. L., "A Taxonomic Problem," Nature, v.140 pp. 1040-1042, 1937.

7. Gilmore, J. S. L., The New Systematics, Claredon, Oxford, 1940.

8. Gower, J. C., "A Comparison of Some Methods of Cluster Analysis," Biometrics, v. 23, pp. 623-637, Dec., 1967.

9. Hadley, G., Nonlinear and Dynamic Programming, Addison-Wesley, Palo Alto, 1964.

10. Hotelling, H., "The Relations Between Two Sets of Variates, Biometrika, v. 28, p. 328, 1936.

11. Rao, C. R., Advanced Statistical Methods in Biometric Research, Wiley, New York, 1952.

12. Rogers, D. J. and Tanimoto, T. T., "A Computer Program for Classifying Plants, Science, v. 132, pp. 1115-1118, 1960.

13. Simpson, G. G., Principles of Animal Taxonomy, Columbia, New York, 1961.

14. Sneath, P. H. A., and Sokal, R. R., "Numerical Taxonomy," Nature, v. 193, pp. 855-860, 1962.

15. Sokal, R. R., "Numerical Taxonomy," Science American, v. 215, pp. 106-116, Dec., 1966.

16. Sokal, R. R. and Sneath, P. H. A., Principles of Numerical Taxonomy, W. H. Frieman Co., San Francisco, 1963.

17. Spranger, E., Types of Men, Stechert-Hafner, Inc., New York

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst. Prof. G. A. Tuck, Code 55 Tk (Adviser) Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1
4.	LT (junior grade) William M. Cima 344 Ridge Avenue Pittsburgh, Pennsylvania 15221	1
5.	Prof. J. R. Borsting, Code 55 Bg Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

A Method of Cluster Analysis

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Master's Thesis; June 1970

5. AUTHOR(S) (First name, middle initial, last name)

William M. Cima

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1970	76	17

8a. CONTRACT OR GRANT NO.

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

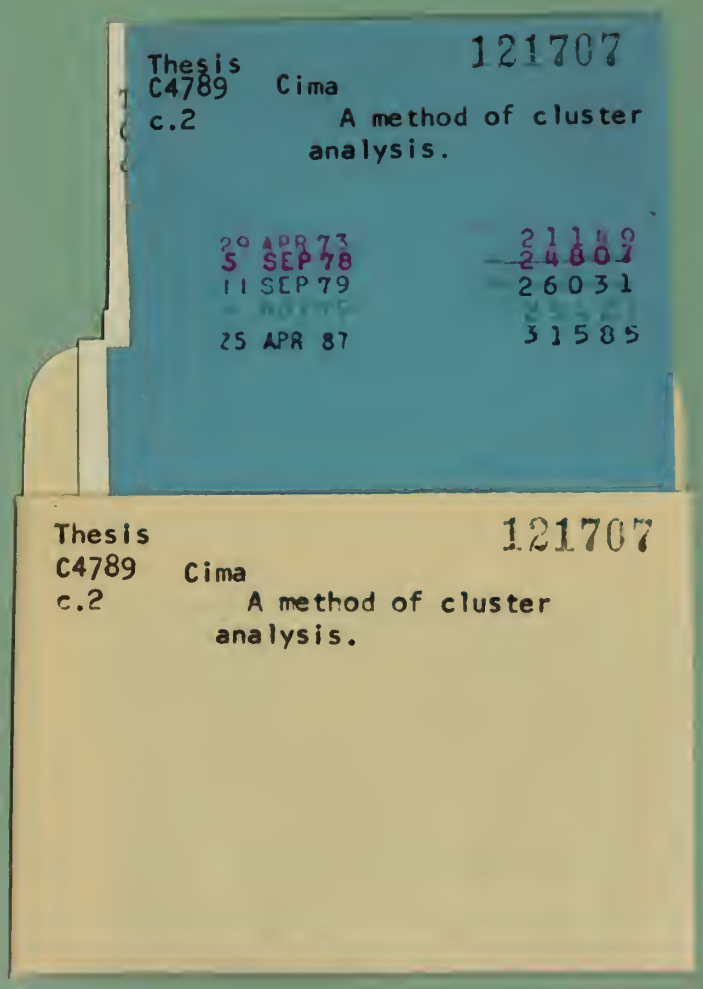
12. SPONSORING MILITARY ACTIVITY

Naval Postgraduate School
Monterey, California 93940

13. ABSTRACT

A method of cluster analysis is presented in which points in n-dimensional space are analyzed through a subdivisive procedure. The points are orthogonally projected onto that line which maximizes their variance and the resulting point distribution is then analyzed with the use of a histogram. Wherever possible, divisions between conglomerates of points are made and each separate clump is subsequently analyzed. Ultimately adjacent groups are combined and analyzed through an analogous technique in an effort to re-unite any points which may have inadvertently deviated from the group with which they truly associate. The method is later refined to allow the detection of groups in several point dispersions which would have appeared as a single conglomeration under the original method. An example is given to illustrate the applicability of the procedure.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Maximum variance line						
Relative maximum variance line						
Conglomerates						
Orthogonal projection						
Cluster analysis						
Statistics						



Thesis
C4789 Cima
c.2 A method of cluster analysis.
121707